



HAL
open science

Développement d'approches prédictives pour l'amélioration génétique : perspectives d'applications pour contribuer à l'adaptation de la vigne aux contraintes imposées par le changement climatique

Vincent Segura

► To cite this version:

Vincent Segura. Développement d'approches prédictives pour l'amélioration génétique : perspectives d'applications pour contribuer à l'adaptation de la vigne aux contraintes imposées par le changement climatique. Génétique des plantes. Université de Montpellier, 2022. tel-04221456

HAL Id: tel-04221456

<https://hal.inrae.fr/tel-04221456>

Submitted on 28 Sep 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0
International License



École Doctorale GAIA
Biodiversité, Agriculture, Alimentation
Environnement, Terre, Eau

Mémoire pour l'obtention de
L'HABILITATION À DIRIGER LES RECHERCHES
de l'Université de Montpellier

**Développement d'approches prédictives pour
l'amélioration génétique : perspectives d'applications
pour contribuer à l'adaptation de la vigne aux
contraintes imposées par le changement climatique**

Vincent SEGURA

UMR AGAP Institut
Univ Montpellier, CIRAD, INRAE, Institut Agro
soutenu le 22 juin 2022

Devant le jury composé de :

<i>Rapporteurs :</i>	Christine DILLMANN	- Professeure, Université Paris-Saclay
	Nathalie OLLAT	- Ingénieure de Recherche, INRAE
	Christèle ROBERT-GRANIÉ	- Directrice de Recherche, INRAE
<i>Examineurs :</i>	Benoît BERTRAND	- Cadre Scientifique, CIRAD
	Mathilde CAUSSE	- Directrice de Recherche, INRAE
	Denis VILE	- Directeur de Recherche, INRAE



INRAE



Amélioration génétique et adaptation des
plantes méditerranéennes et tropicales

Avant-propos & remerciements

Ça y est c'est mon tour, après avoir repoussé l'échéance, et cela malgré les encouragements reçus lors de mes précédentes évaluations, j'ai enfin su trouver les ressources et la motivation nécessaires pour réaliser cet exercice. Il faut dire que ce n'est pas facile d'y allouer du temps, puisqu'on arrive toujours à trouver quelque chose de plus prioritaire à faire... En tout cas, je suis bien content d'avoir réussi à le mettre tout en haut de ma liste de tâches, et j'en profite d'ailleurs pour m'excuser auprès de mes collègues d'avoir pris un peu de retard sur certains dossiers ces dernières semaines. A vrai dire, ça valait quand même le coût de se poser un peu, de sortir la tête du guidon, pour faire un bilan général et réfléchir à la façon dont s'agentent et s'organisent les différents travaux et projets mis en œuvre au quotidien.

Plus de 15 ans après mes débuts dans le monde de la recherche, je me retrouve finalement quasiment au même endroit, mais entre temps j'ai "grandi" et il s'est passé plein de choses. J'ai notamment eu la chance de rencontrer et de travailler avec de nombreuses personnes que je souhaiterais ici remercier. Je voudrais d'abord mentionner les personnes qui m'ont encadré et qui m'ont accordé leur confiance, par ordre chronologique : Dominique, Évelyne, Anne, Magnus, Christophe, Catherine, Patrice. Je voudrais aussi remercier chaleureusement toutes les personnes avec qui au quotidien j'ai pu travailler, qu'il s'agisse de personnels technique, administratif, ou scientifique des différentes équipes ou unités dans lesquelles j'ai pu travailler. Sans toutes ces personnes bien sûr je n'en serai pas à rédiger ce document. Il s'agit principalement des personnels de l'équipe AFEF à Montpellier, du projet Artemisia à York, du groupe "Nordborg" à Vienne, des unités AGPF/BioForA et GBFor à Orléans, et de l'équipe DAAV et de l'UMR AGAP à Montpellier. Il y aurait beaucoup trop de personnes à citer ici, et j'ai vraiment peur d'en oublier... Je m'en tient donc à quelques mentions spéciales pour chacune des différentes périodes de ma vie dans la recherche scientifique. Merci à Alex avec qui, malgré la très grande distance, nous n'avons cessé de garder un contact à la fois scientifique et amical. Merci à Loïc de m'avoir informé de l'ouverture du poste à la mobilité, c'est génial de pouvoir se retrouver dans le même bureau 15 ans après, avec certaines constantes : un grand tableau blanc, une étuve/bar et plein d'idées. Merci aux amis rencontrés lors des séjours à l'étranger, ça laisse des souvenirs impérissables, à la fois au boulot mais aussi à l'extérieur : Tomasz et Agnieszka, les deux T(h)eresa, Arthur, Bjarni, Ümit, Envel. Merci aux collègues de course Orléanais, ces sorties méridiennes à discuter science tout en galopant me manquent, et à ceux qui ont fait que ces 10 années en région Centre soient particulièrement agréables et enrichissantes. Merci aussi à mes nouveaux collègues de l'équipe DAAV et de l'UMR AGAP pour leur accueil, malgré des conditions assez difficiles ces dernières années. Merci à tous les étudiants en stage, thèse et aux post-docs que j'ai eu la chance d'encadrer ou avec qui j'ai pu collaborer, mention spéciale à Aurélien, Thibaud, Marie, Charlotte et Abdou, ainsi qu'à Véro, Catherine, Leo, Jean-Paul, Odile, Harold, Agnès, Loïc, Patrice pour les co-encadrements. Merci aux collègues avec qui j'ai pu et/ou je compte encore

collaborer et plus particulièrement Christopher, Stéphane, Régis, Renaud, Thierry, Charles, Martin, Tristan, Aude.

Enfin, je voudrai remercier mes proches et notamment Mélanie mon épouse pour m'avoir suivi dans toutes ces aventures, merci pour ta capacité d'adaptation et ton soutien sans faille. Merci également à nos trois filles qui malgré leur jeune âge ont gentiment accepté de me laisser tranquille ces dernières semaines pour que je puisse rédiger ce document.

Texte relatif à l'intégrité scientifique

Je déclare avoir respecté, dans la conception et la rédaction de ce mémoire d'HDR, les valeurs et principes d'intégrité scientifique destinés à garantir le caractère honnête et scientifiquement rigoureux de tout travail de recherche, visés à l'article L.211-2 du Code de la recherche et énoncés par la Charte nationale de déontologie des métiers de la recherche et la Charte d'intégrité scientifique de l'Université de Montpellier. Je m'engage à les promouvoir dans le cadre de mes activités futures d'encadrement de recherche.

Table des matières

1	CV étendu	1
1.1	Formation	1
1.2	Expérience, dont stages dans des laboratoires français et étrangers	1
1.3	Contrats de recherche	2
1.4	Enseignement	3
1.5	Encadrement	3
1.5.1	Post-doctorants	3
1.5.2	Doctorants	3
1.5.3	Stagiaires	4
1.6	Expertise scientifique	5
1.6.1	Participation à des jurys de thèse	5
1.6.2	Participation à des comités de thèse	6
1.6.3	Participation à des jurys de recrutement	8
1.6.4	Relecture d'articles scientifiques	8
1.7	Publications	8
1.7.1	Analyse bibliométrique	8
1.7.2	Liste des publications	9
2	Rapport sur les travaux de recherche	15
2.1	Étude des déterminismes génétiques de caractères complexes par des approches de détection de QTLs	15
2.1.1	Déterminismes génétiques de l'architecture aérienne chez le pommier	15
2.1.2	Génétique et amélioration d' <i>Artemisia annua</i> L. pour une production durable d'antipaludiques à base d'artémisinine	16
2.1.3	Bilan	17
2.2	Développements méthodologiques pour la GWAS et collaborations associées	18
2.2.1	Le modèle linéaire mixte pour la génétique d'association	18
2.2.2	MLMM	20
2.2.3	MTMM	23
2.2.4	Collaborations associées	26
2.3	Biologie intégrative de la production de biomasse chez le peuplier	26

2.3.1	Contexte	26
2.3.2	Questions de recherche et plan de travail	27
2.3.3	Principaux résultats	30
2.3.4	Au-delà du transcriptome, implication de la méthylation de l'ADN dans la variabilité phénotypique	40
2.4	Prédiction phénotypique	40
2.4.1	Définition et preuve de concept	41
2.4.2	Application chez la vigne	44
2.5	Conclusions sur la partie bilan	46
3	Perspectives	49
3.1	Contexte, enjeux et objectifs	50
3.2	Architecture génétique de caractères d'intérêt en réponse à la contrainte hydrique	53
3.2.1	Matériel d'étude : un panel de diversité génétique	53
3.2.2	Le projet G2WAS	54
3.2.3	Fonctionnement végétatif	54
3.2.4	Métabolisme de la baie	57
3.2.5	A plus long terme : validations au vignoble et analyses de diversité entre génotypes extrêmes	60
3.3	Outils d'aide à la sélection	61
3.3.1	Vers l'utilisation des prédictions génomique et phénotypique dans les programmes d'amélioration génétique de la vigne	61
3.3.2	Les cas particulier des croisements d'intégration	64
3.4	Conclusions sur la partie perspectives	66
	Bibliographie	67
	Annexes	77

Table des figures

1	Analyse bibliométrique	8
2	Résultats des simulations effectuées pour évaluer les performances de la méthode MLMM	22
3	Résultats des simulations effectuées pour évaluer les performances de la méthode MTMM	25
4	Illustration de la démarche mise en œuvre dans le cadre du projet SYBIOPOP	28
5	Représentation graphique de l'origine géographique et de la structure de la diversité génétique de la collection de peupliers noirs	29
6	Illustration de modèles de calibration NIRS développés pour le phénotypage à haut-débit des propriétés du bois	30
7	Génétique d'association à l'échelle du transcriptome pour la circonférence mesurée à Savigliano	33
8	Indices de différenciation entre populations de peuplier noirs	34
9	Effet du top SNP sur la partition de variance inter-sites	35
10	Illustration des associations à l'échelle populationnelle	35
11	Illustration de la différenciation génétique observée pour le top SNP	36
12	Carte des eQTLs détectés	38
13	Relation entre changement d'importance des prédicteurs et avantage prédictif en multi-omiques	39
14	Précisions des prédictions phénotypique et génomique chez le blé et le peuplier	43
15	Précisions des prédictions phénotypique et génomique chez la vigne	46
16	Organigramme de l'équipe DAAV	49
17	Schéma de sélection de la vigne en France	50
18	Impacts du changement climatique sur la vigne et le vin	51
19	Résultats des calibrations NIRS pour des caractères fonctionnels chez la vigne.	56
20	Stades de développement d'une baie	58
21	Dynamique du programme d'amélioration génétique de la vigne en France	64

Chapitre 1

CV étendu

VINCENT SEGURA

Né le 06/06/1981 à Montpellier (34)
Marié, 3 enfants

Chargé de Recherche de Classe Normale, INRAE
Equipe Diversité, Adaptation et Amélioration de la Vigne (DAAV)
UMR AGAP Institut, Univ. Montpellier, CIRAD, INRAE, Institut Agro
F-34398 Montpellier
vincent.segura@inrae.fr

1.1 Formation

- 2004-2007 : **Doctorat**, *Biologie fonctionnelle des Plantes*, Montpellier SupA-gro.
- 2003-2004 : **DEA**, *Développement et adaptation des plantes, biologie moléculaire intégrative*, Université de Montpellier 2.
- 2000-2003 : **IUP**, *Biologie appliquée aux productions végétales et aux industries agro-alimentaires*, Université de Picardie Jules Verne.
- 1998-2000 : **DEUG**, *Sciences de la Vie option Sciences Biologiques Naturelles*, Université de Montpellier 2.
- 1998 : **Baccalauréat Général**, *Série Scientifique*, lycée Joffre, Montpellier.

1.2 Expérience, dont stages dans des laboratoires français et étrangers

- 2019- : **Chargé de Recherche INRAE**, *Analyse de la diversité génétique de la vigne à l'échelle du génome entier et exploitation en sélection*, Équipe

DAAV, UMR AGAP Institut, Univ. Montpellier, CIRAD, INRAE, Institut Agro, Montpellier.

- 2009-2019 : **Chargé de Recherche INRAE**, *Biologie intégrative de la production de biomasse chez le peuplier*, Équipe Prédiction, UMR BioForA (anciennement AGPF), INRAE, ONF, Orléans.
- 2010-2011 : **Post-doctorant**, *Développements méthodologiques pour les études d'association pangénomiques : modèles multi-locus et multi-caractères*, Équipe génétique des populations, Gregor Mendel Institute for Molecular Plant Biology, Vienne, Autriche. Responsable scientifique : Magnus Nordborg.
- 2008-2009 : **Post-doctorant**, *Génétique et amélioration d'Artemisia annua pour la production d'une molécule antipaludique, l'artémisinine*, Équipe génétique et amélioration des plantes, Projet Artemisia, Université de York, Royaume-Uni. Responsable scientifique : Anne Rae.
- 2004-2007 : **Doctorant**, *Étude des déterminismes génétiques de l'architecture du pommier*, UMR Développement et Amélioration des Plantes, Équipe Architecture et Fonctionnement des Espèces Fruitières (AFEF), INRA-Montpellier SupAgro, Directrice de thèse : Evelyne Costes.
- 2004 : **Stagiaire (DEA)**, *Étude de l'héritabilité et des corrélations génétiques de caractères architecturaux chez le pommier*, UMR Biologie du Développement des Plantes Pérennes Cultivées (BDPPC), Équipe AFEF, INRA-Montpellier SupAgro, Responsable Scientifique : Evelyne Costes.
- 2003 : **Stagiaire (Maîtrise)**, *Étude de l'héritabilité de caractères architecturaux chez une descendance d'abricotiers*, UMR BDPPC, Équipe AFEF, INRA-Montpellier SupAgro, Responsable Scientifique : Dominique Fournier.

1.3 Contrats de recherche

- 2021-2024, **OASIs (CASDAR)**, *Outils d'Aide à la décision pour accélérer la Sélection dans les croisements d'Intégration chez la vigne*, Coordinateur du projet.
- 2020-2023, **G2WAS (ANR-19-CE20-0024)**, *Architecture génétique de la tolérance au stress hydrique chez la vigne*, Co-coordonateur du WP4.
- 2020-2023, **SelGenVit (ANR-19-ECOM-0006)**, *Sélection génomique au service de l'amélioration de la vigne pour la diversification et le déploiement de variétés résistantes à forts potentiels œnologiques*, Coordinateur du WP2.
- 2018-2022, **EPITREE (ANR-17-CE32-0009)**, *Impacts évolutifs et fonctionnels des variations épigénétiques chez les arbres forestiers*, Coordinateur du WP2.
- 2018-2021, **EPINET (INRA-Selgen)**, *Prédiction de valeurs génétiques guidée par des réseaux de gènes et à travers une utilisation explicite de l'épistasie*, Coordinateur du projet.

- 2015-2019, **TOPWOOD** (H2020-MSCA-RISE-2014), *Outils de phénotypage du bois : propriétés, fonctions et qualité*, Coordinateur du WP3.
- 2014-2018, **SYBIOPOP** (ANR-13-JSV6-0001), *Une approche de biologie intégrative pour améliorer le peuplier en vue de sa valorisation en bio-raffinerie grâce à une meilleure compréhension de l'architecture génétique de la production et de la qualité de la biomasse lignocellulosique*, Coordinateur du projet.

1.4 Enseignement

- 2021-, **Biologie intégrative appliquée à l'étude du déterminisme génétique de la production de bois chez le peuplier**, *Master Biologie, Agrosciences*, Université de Montpellier (1,5 heures de cours / an).
- 2020-, **Génétique d'association et prédiction génomique**, *Ingénieur agronome et Master 3A*, Institut Agro | Montpellier SupAgro (10,5 heures de cours et 3,5 heures de TD / an).
- 2014-2019, **Biostatistiques**, *Master Agrosciences, Environnement, Territoire, Paysage, Forêt*, Université d'Orléans (2 fois 11 heures / an).
- 2012-2019, **Logiciel R**, *INRAE Val-de-Loire* (6 sessions sur la période)
- 2012, **Etudes d'associations pan-génomiques avec des modèles mixtes**, *Ecole d'été EVOLTREE*, Uppsala, Suède, 5-7 septembre 2012. (21 heures en collaboration avec A. Hancock et B.J. Vilhjalmsen).

1.5 Encadrement

1.5.1 Post-doctorants

- 2016-2018 : **Aurélien Chateigner**, *Génétique quantitative de la variation transcriptomique chez le peuplier noir* (co-encadrement avec Leopoldo Sanchez).
- 2013-2015 : **Nassim Belmokhtar**, *Prédiction et analyse de la variabilité génétique de la conversion de la biomasse en bioéthanol 2G chez le miscanthus et le peuplier* (co-encadrement avec Jean-Paul Charpentier).

1.5.2 Doctorants

- 2021- : **Flora Tavernier**, *Metab'EAU : variabilité génétique du métabolome du raisin en réponse à une contrainte hydrique*, Institut Agro | Montpellier SupAgro (co-encadrement avec Charles Romieu).
- 2019- : **Abdou Rahmane Wade**, *Amélioration de la précision de prédiction pan-génomique du phénotype par une approche de biologie des systèmes*, AgroParisTech (co-encadrement avec Harold Duruflé et Leopoldo Sanchez).

- 2019-2021 : **Charlotte Brault**, *Optimisation de l'amélioration génétique chez la vigne avec les prédictions génomique et phénotypique*, Institut Agro | Montpellier SupAgro (co-encadrement avec Agnès Doligez, Loïc Le Cunff et Patrice This).
- 2014-2017 : **Mesfin Nigussie Gebreselassie**, *Architecture génétique de la production et de la qualité de la biomasse chez le peuplier noir pour une valorisation en bio-raffinerie*, Université d'Orléans (co-encadrement avec Catherine Bastien).
- 2009-2013 : **Redouane El Malki**, *Architecture génétique de caractères cibles pour la culture du peuplier en taillis à courte rotation*, Université d'Orléans (co-encadrement avec Véronique Jorge et Catherine Bastien).

1.5.3 Stagiaires

- 2021 : **Theresa Herbold**, *Variabilité phénotypique des baies de raisin basée sur la spectroscopie dans le proche infrarouge (SPIR) dans un plan de croisement demi-diallèle*, Master 2, Plant Science, Université d'Hohenheim.
- 2021 : **Virgilio Freitas**, *Phénotypage à haut-débit de caractères fonctionnels chez la vigne en vue de caractériser la variabilité génétique de sa réponse aux contraintes liées au changement climatique*, Master 1, 3A, parcours sélection et évolution des plantes méditerranéennes et tropicales, Institut Agro | Montpellier SupAgro (co-encadrement avec Aude Coupel-Ledru).
- 2021 : **Miguel Thomas**, *Preuve de concept de l'utilisation de données phénotypiques comme outil d'aide à la sélection dans un plan de croisement demi-diallèle chez la vigne*, élève Ingénieur INP-ENSAT spécialité Agrobiosciences Végétales (co-encadrement avec Loïc Le Cunff).
- 2020 : **Juliette Lazerges**, *Prédictions génomique et phénotypique interpolation chez la vigne*, élève Ingénieur INP-ENSAT spécialité Agrobiosciences Végétales (co-encadrement avec Charlotte Brault, Agnès Doligez et Loïc Le Cunff).
- 2020 : **Manon Malestroit**, *Analyse statistique de données de caractérisation des ressources génétiques de la vigne pour les principaux bioagresseurs*, Master 2, Statistiques pour les sciences de la vie, Univ. Montpellier (co-encadrement avec Agnès Doligez, Jean-Pierre Péros et Thierry Lacombe).
- 2019 : **Ilyas Bouhaba**, *Estimation d'effets génétiques indirects (de compétition) et de leurs déterminismes sous-jacents sur la production de biomasse chez le peuplier noir*, Master 2, 3A, parcours sélection et évolution des plantes méditerranéennes et tropicales, Montpellier SupAgro (co-encadrement avec Leopoldo Sanchez).
- 2018 : **Gaëlle Desclos**, *Étude de la variabilité génétique de la densité du bois chez le peuplier noir*, Master 1, Biologie des Organismes, Populations et Écosystèmes, Université d'Orléans.

- 2016 : **Souhila Amanzougarene**, *Étude de la diversité nucléotidique de séquences transcrites chez le peuplier noir*, Master 2, Bioinformatique et Génomique, Université de Rennes 1 (co-encadrement avec Odile Rogier et Véronique Jorge).
- 2015 : **Vincent Simonot**, *Développement d'une méthode de phénotypage haut débit par spectrométrie proche infrarouge pour la détermination de la durabilité naturelle chez le mélèze*, Master 2, Biologie des Organismes, Populations et Écosystèmes, Université d'Orléans (co-encadrement avec Luc Pâques et Jean-Paul Charpentier).
- 2013 : **Siham Balit**, *Étude de la variabilité de l'aptitude à former du bois de tension chez 10 clones de peuplier*, Master 1, Biologie des Organismes, Populations et Écosystèmes, Université d'Orléans (co-encadrement avec Jean-Paul Charpentier).
- 2012 : **Violaine Murciano**, *Construction de modèles de prédiction des propriétés du bois de peuplier à partir de données de spectrométrie en moyen-infrarouge*, Master 1, Statistiques et Recherche Opérationnelle, Université d'Orléans.
- 2010 : **Justine Guet**, *Étude de la variabilité de l'aptitude à former du bois de tension chez différents génotypes de peuplier*, Master 1, Écosystèmes Terrestres, Université d'Orléans (co-encadrement avec Gilles Pilate et Françoise Laurans).
- 2006 : **Randa Hammou-Ahabchane**, *Identification de prédicteurs des premières floraisons dans une descendance de pommiers au moyen de modèles linéaires*, Master 2, Génie Statistique et Informatique, Université d'Aix-Marseille 1 (co-encadrement avec Evelyne Costes).
- 2005 : **Mayeul Milien**, *Héritabilité et corrélations génétiques des caractères architecturaux dans une descendance de pommiers*, Master 1, Biologie Fonctionnelle des Plantes, Université de Montpellier 2 (co-encadrement avec Evelyne Costes).

1.6 Expertise scientifique

1.6.1 Participation à des jurys de thèse

- 2020, **Anthony Bernard**, *Étude des ressources génétiques Juglans en vue de la mise en place d'une sélection assistée par marqueurs*, Université de Bordeaux.
- 2020, **Romuald Laso-Jadart**, *Epipelagic marine plankton through the lens of population genomics : a molecular study using metagenomic and metatranscriptomic data of Tara Oceans expeditions*, Université Paris-Saclay.
- 2019, **Mamadou Dia Sow**, *Rôle fonctionnel de l'épigénétique (Méthylation de l'ADN) dans la réponse du peuplier à des variations de disponibilité en eau du sol*, Université d'Orléans.

- 2019, **Jiantao Zhao**, *Combining association and haplotype studies towards the improvement of fruit quality in tomato*, Université d'Avignon.
- 2019, **Simon Rio**, *Contributions to Genomic Selection and Association mapping in structured and admixed populations : application to maize*, Université Paris-Saclay.
- 2018, **Florent Guinot**, *Statistical learning for omics association and interaction studies based on blockwise feature compression*, Université Paris-Saclay.
- 2016, **Hélène Lagraulet**, *Plasticité phénotypique et architecture génétique de la croissance et de la densité du bois du pin maritime (*Pinus pinaster* Ait.)*, Université de Bordeaux.
- 2014, **Guillaume Bauchet**, *Génétique d'association chez la tomate – Une stratégie pour identifier des locus importants pour la qualité du fruit*, Université d'Avignon.

1.6.2 Participation à des comités de thèse

- 2021- : **Laila Aqbouch**, *Genomics and adaptation of cultivated olive to climate change*, Institut Agro | Montpellier SupAgro.
- 2020- : **Mathilde Milan**, *Orientation des feuilles chez la vigne : variabilité génétique et conséquences sur la sensibilité aux températures élevées et sur l'efficacité d'utilisation de l'eau*, Université de Montpellier.
- 2019- : **Louis Blois**, *Etude de l'architecture génétique des réponses au déficit hydrique chez *Vitis berlandieri**, Université de Bordeaux.
- 2019- : **Pauline Robert**, *Evaluation et mise en œuvre de la sélection phénotypique en pré-breeding et sélection du blé tendre*, Université de Clermont-Auvergne.
- 2019- : **Severine Monnot**, *Sélection pour la résistance oligogénique à plusieurs virus chez le concombre : de l'implémentation de méthodes de sélection génomique à l'étude des répertoires de gènes impliqués dans la résistance*, Université d'Avignon.
- 2019- : **Estelle Bineau**, *Génétique d'association chez la tomate pour identifier des locus, des gènes et des allèles importants dans la sélection de variétés de tomate de qualité*, Université d'Avignon.
- 2018-2021 : **Alexandre Mallet**, *S'affranchir de l'effet de l'eau pour permettre une caractérisation spectroscopique des déchets organiques humides*, Université de Montpellier.
- 2017-2021 : **Mariem Nsibi**, *GWAS and genomic selection in apricot*, Université de Montpellier.
- 2017-2020 : **Romuald Laso-Jadart**, *Linking Allele-specific expression and natural selection in *O. similis* populations of Artic Seas*, Université Paris-Saclay.

- 2016-2019 : **Clément Mabire**, *Étude de l'effet des variations structurales de type présence/absence sur les caractères d'intérêt agronomique et l'hétérosis au moyen d'une puce de génotypage haut débit et par des approches d'association et de prédiction génomique chez le maïs*, Université Paris-Saclay.
- 2015-2018 : **Agathe Maupetit**, *Potentiel évolutif et déterminisme génétique de caractères d'agressivité et morphologiques de l'agent de la rouille du peuplier, *Melampsora larici-populina**, Université de Lorraine.
- 2015-2018 : **Marie Pégard**, *New models for implementation of genome-wide evaluation in black poplar breeding program*, Université d'Orléans.
- 2014-2017 : **Marie Garavillon-Tournaire**, *Étude de la variabilité génétique des réponses écophysiological et moléculaire associées au transport d'eau dans dans la feuille de peuplier noir en carence hydrique*, Université de Clermont-Auvergne.
- 2013-2017 : **Élise Albert**, *Déterminants génétiques de la réponse au déficit hydrique chez la tomate par des approches de génétique de liaison, d'association et de transcriptomique*, Université d'Avignon.
- 2013-2017 : **Anne-Laure Le Gac**, *Méthylation de l'ADN et plasticité phénotypique en réponse à des variations de disponibilité en eau chez le peuplier*, Université d'Orléans.
- 2012-2016 : **Héloïse Giraud**, *Genetic analysis of hybrid value for silage maize in multiparental designs : QTL detection and genomic selection*, Université Paris-Saclay.
- 2012-2016 : **Alix Allard**, *Estimation de la valeur génétique de pommiers hybrides pour plusieurs caractères génétiques d'intérêt agronomique et sur la base de leur apparentement*, Montpellier SupAgro.
- 2012-2015 : **Justine Guet**, *Expression d'une variabilité génétique pour la phénologie de croissance, l'efficacité d'utilisation de l'eau et la résistance à la cavitation chez le peuplier noir (*Populus nigra L.*)*, Université d'Orléans.
- 2011-2015 : **Diane Leforestier**, *Localisation de régions du génome du pommier contrôlant la variation de caractères de qualité du fruit et de résistance aux maladies : signatures de sélection et génétique d'association*, Université d'Angers.
- 2010-2013 : **Agota Fodor**, *La sélection génomique appliquée chez la vigne, évaluation et utilisation*, Montpellier SupAgro.
- 2009-2013 : **Ines Ben Sadok**, *Étude du déterminisme génétique de l'architecture de l'olivier en vue de l'amélioration de la production et de la qualité du fruit*, Montpellier SupAgro.

1.6.3 Participation à des jurys de recrutement

- 2020 : concours CRCN, prédiction et sélection génomiques
- 2020 : concours CRCN, génétique en aquaculture
- 2019 : concours IR, ingénieur-e biologiste en laboratoire

1.6.4 Relecture d'articles scientifiques

Depuis 2011, j'ai effectué la relecture de **36 articles** pour les journaux suivants : *Annals of Forest Sciences* (1), *Bioenergy Research* (2), *BMC Bioinformatics* (1), *BMC Genomics* (1), *BMC Plant Biology* (1), *Evolutionary Applications* (1), *Frontiers in Plant Science* (1), *Heredity* (3), *Molecular Ecology* (2), *Nature Communications* (7), *New Phytologist* (3), *Plant Journal* (1), *Plant Methods* (1), *Sylvae Genetica* (1), *Theoretical and Applied Genetics* (2), *Tree Genetics and Genomes* (5), *Tree Physiology* (3).

1.7 Publications

1.7.1 Analyse bibliométrique

Web of Science (données collectées le 24/11/2021) : 35 publications, 1648 citations, indice h : 20 (FIGURE 1).

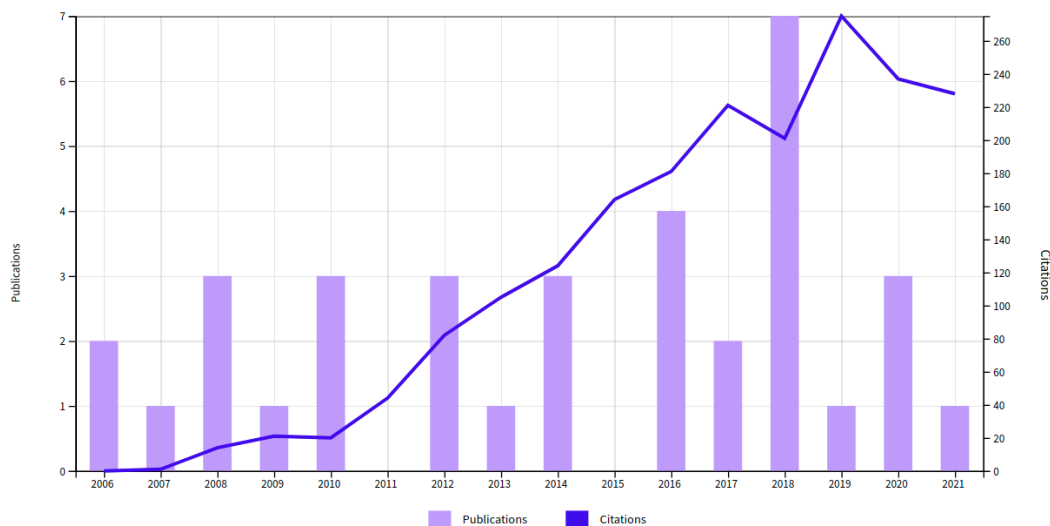


FIGURE 1 – **Analyse bibliométrique**

Analyse réalisée à partir des données collectées le 24/11/2021 sur le Web of Science.

1.7.2 Liste des publications

Les noms des stagiaires, doctorants ou post-doctorants co-encadrés sont soulignés.

Articles soumis

1. Wade AR, Durufle H, Sanchez L, **Segura V**. eQTLs are key players in the integration of genomic and transcriptomic data for phenotype prediction. *Soumis à BMC Genomics*. doi: 10.1101/2021.09.07.459279.
2. Brault C, **Segura V**, This P, Le Cunff L, Flutre T, François P, Pons T, Péros J-P, Doligez A. Across-population genomic prediction in grapevine opens up promising prospects for breeding. *Soumis à Horticulture Research*. doi: 10.1101/2021.07.29.454290.
3. Persoons A, Maupetit A, Louet C, Andrieux A, Lipzen A, Barry KW, Na H, Adam C, Grigoriev IV, **Segura V**, Duplessis S, Frey P, Halkett F, De Mita S. Genomic signatures of a major adaptive event in the pathogenic fungus *Melampsora larici-populina*. *Soumis à Genome Biology and Evolution*. doi: 10.1101/2021.04.09.439223.
4. Coupel Ledru A, Pallas B, Delalande M, **Segura V**, Guitton B, Muranty H, Durel C-E, Regnard J-L, Costes E. Tree architecture, light interception and water use related traits are controlled by different genomic regions in an apple tree core collection. *Soumis à New Phytologist*.

Articles sous-presse

5. Robert P, Brault C, Rincent R, **Segura V**. Phenomic selection : a new and efficient alternative to genomic selection. In : *Genomic prediction of complex traits*, Springer Nature.

2021

6. Sow MD, Le Gac A-L, Fichot R, Lanciano S, Delaunay A, Le Jan I, Lesage-Descauses M-C, Citerne S, Caius J, Brunaud V, Soubigou-Taconnat L, Cochard H, **Segura V**, Chaparro C, Grunau C, Daviaud C, Tost J, Brignolas F, Strauss SH, Mirouze M, Maury S. RNAi suppression of DNA methylation affects the drought stress response and genome integrity in transgenic poplar. *New Phytologist*, 232, 80-97, doi: 10.1111/nph.17555

2020

7. Pégard M, **Segura V**, Muñoz F, Bastien C, Jorge V, Sanchez L. Favorable conditions for genomic evaluation to outperform classical pedigree evaluation highlighted by a proof-of-concept study in poplar. *Frontiers in Plant Science*, 11, 1552. doi: 10.3389/fpls.2020.581954

8. Chateigner A, Lesage-Descauses M-C, Rogier O, Jorge V, Leplé J-C, Brunaud V, Paysant-Le Roux C, Soubigou-Taconnat L, Martin-Magniette M-L, Sanchez L, **Segura V**. Gene expression predictions and networks in natural populations supports the omnigenic theory. *BMC Genomics*, 21, 416. doi: 10.1186/s12864-020-06809-2
9. Sergent A-S, **Segura V**, Charpentier J-P, Dalla-Salda G, Fernández M-E, Rozenberg P, Martinez-Meier A. Assessment of resistance to xylem cavitation in cordilleran cypress using near-infrared spectroscopy. *Forest Ecology and Management*, 462, 117943. doi: 10.1016/j.foreco.2020.117943

2019

10. Chauvin T, Cochard H, **Segura V**, Rozenberg P. Native-source climate determines the Douglas-fir potential of adaptation to drought. *Forest Ecology and Management*, 444, 9-20. doi: 10.1016/j.foreco.2019.03.054

2018

11. Rincent R, Charpentier J-P, Faivre-Rampant P, Paux E, Le Gouis J, Bastien C, **Segura V**. Phenomic selection is a low-cost and high-throughput method based on indirect predictions : proof of concept on wheat and poplar. *G3 : Genes, Genomes, Genetics*, 8 (12), 3961-3972. doi: 10.1534/g3.118.200760
12. Rogier O, Chateigner A, Amanzougarene S, Lesage-Descauses M-C, Balzergue S, Brunaud V, Caius J, Soubigou-Taconnat L, Jorge V, **Segura V**. Accuracy of RNAseq based SNP discovery and genotyping in *Populus nigra*. *BMC Genomics*, 19, 909. doi: 10.1186/s12864-018-5239-z
13. Sow MD, **Segura V**, Chamailard S, Jorge V, Delaunay A, Lafon-Placette C, Fihcot R, Faivre-Rampant P, Villar M, Brignolas F, Maury S. Narrow-sense heritability and Pst estimates of DNA methylation in three *Populus nigra* L. populations under contrasting water availability. *Tree Genetics & Genomes*, 14, 78. doi: 10.1007/s11295-018-1293-6
14. Le Gac A-L, Lafon-Placette C, Chauveau D, **Segura V**, Delaunay A, Fichot R, Marron M, Le Jan I, Berthelot A, Bodineau G, Bastien J-C, Brignolas F, Maury S. Winter-dormant shoot apical meristem in poplar trees shows environmental epigenetic memory, *Journal of Experimental Botany*, 69 (20), 4821-4837. doi: 10.1093/jxb/ery271
15. Lafon-Placette C, Le Gac A-L, Chauveau D, **Segura V**, Delaunay A, Lesage-Descauses M-C, Hummel I, Cohen D, Jesson B, Le Thiec D, Bogeat-Triboulot M-B, Brignolas F, Maury S. Changes in the epigenome and transcriptome of the poplar shoot apical meristem in response to water availability affect preferentially hormone pathways. *Journal of Experimental Botany*, 69 (3), 537-551. doi: 10.1093/jxb/erx409
16. Sow MD, Allona I, Ambroise C, Conde D, Fichot R, Gribkova S, Jorge V, Le-Provost G, Pâques L, Plomion C, Salse J, Sanchez-Rodriguez L, **Segura**

V, Tost J, Maury S. Epigenetics in forest trees : state of the art and potential implications for breeding and management in a context of climate change. In *Advances in Botanical Research*, 88, 387-453. Academic Press Inc. doi: 10.1016/bs.abr.2018.09.003

17. Chaix G, Giordanengo T, **Segura V**, Mourey N, Charrier B, Charpentier J-P. Near infrared spectroscopy, a new tool to characterize wood for use by the cooperage industry. In : *Chemistry of lignocellulosics : current trends*, 42-65. Boca Raton : CRC Press. doi: 10.1201/b20936

2017

18. Gebreselassie MN, Ader K, Boizot N, Millier F, Charpentier J-P, Alves A, Simoes R, Rodrigues JC, Bodineau G, Fabbrini F, Sabatti M, Bastien C, **Segura V**. Near-infrared spectroscopy enables the genetic analysis of chemical properties in a large set of wood samples from *Populus nigra* (L.) natural populations. *Industrial Crops and Products*, 107, 159-171. doi: 10.1016/j.indcrop.2017.05.013
19. Bauchet G, Grenier S, Samson N, **Segura V**, Kende A, Beekwilder J, Cankar K, Gallois J-L, Gricourt J, Bonnet J, Baxter C, Grivet L, Causse M. Identification of major loci and genomic regions controlling acid and volatile content in tomato fruit : implications for flavor improvement. *New Phytologist*, 215, 624-641. doi: 10.1111/nph.14615

2016

20. Albert E, **Segura V**, Gricourt J, Bonnefoi J, Derivot L, Causse M. Association mapping reveals the genetic architecture of tomato response to water deficit : focus on major fruit quality traits. *Journal of Experimental Botany*, 67, 6413-6430. doi: 10.1093/jxb/erw411
21. Pulkka S*, **Segura V***, Harju A, Tapanila T, Tanner J, Pâque LE, Charpentier J-P. Prediction of stilbene content from heartwood increment cores of Scots pine using near infrared spectroscopy methodology. *Journal of NIRS*, 24, 517-528. doi: 10.1255/jnirs.1225
22. Faivre-Rampant P, Zaina G, Jorge V, Giacomello S, **Segura V**, Scalabrin S, Guérin V, De Paoli E, Aluome C, Viger M, Cattonaro F, Payne P, StephenRaj P, Le Paslier MC, Berard A, Allwright M, Villar M, Taylor G, Bastien C, Morgante M. New resources for genetic studies in *Populus nigra* : genome-wide SNP discovery and development of a 12k Infinium array. *Molecular Ecology Resources*, 16, 1023-1036. doi: 10.1111/1755-0998.12513
23. Plomion C, Bastien C, Bogeat-Triboulot M-B, Bouffier L, Déjardin A, Duplessis S, Fady B, Heuertz M, Le Gac A-L, Le Provost G, Legué V, Lelu-Walter M-A, Leplé J-C, Maury S, Morel A, Oddou-Muratorio S, Pilate G,

*. contribution égale

Sanchez L, Scotti I, Scotti-Saintagne C, **Segura V**, Trontin J-F, Vacher V. Forest tree genomics : 10 achievements from the past 10 years and future prospects. *Annals of Forest Science*, 73, 77-103. doi: 10.1007/s13595-015-0488-3

2014

24. Giraud H, Lehermeier C, Bauer C, Falque M, **Segura V**, Bauland C, Camisan C, Campo L, Meyer N, Ranc N, Schipprack W, Flament P, Melchinger AE, Menz M, Moreno-González J, Ouzunova M, Charcosset A, Schön CC, Moreau L. QTL detection for hybrid performance using different allele coding methods in multi-line crosses of maize (*Zea mays* L.) reveals multi-allelic series at different loci in the flint and dent heterotic groups. *Genetics* 198, 1717-1734. doi: 10.1534/genetics.114.169367
25. Fodor A, **Segura V**, Denis M, Neuenschwander S, Fournier-Level A, Chatelet P, Homa FAA, Lacombe T, This P, Le Cunff L. Genome-wide prediction methods in highly diverse and heterozygous species : proof-of-concept through simulation in grapevine. *PLoS ONE*, 9, e110436. doi: 10.1371/journal.pone.0110436
26. Sauvage C, **Segura V**, Bauchet G, Stevens R, Thi Do P, Nikoloski Z, Fernie AR, Causse M. Genome Wide Association in tomato reveals 44 candidate loci for fruit metabolic traits. *Plant Physiology*, 165, 1120-1132. doi: 10.1104/pp.114.241521

2013

27. Townsend T, **Segura V**, Chigeza G, Penfield T, Rae AM, Harvey D, Bowles DJ, Graham IA. The use of combining ability analysis to identify elite parents for *Artemisia annua* F1 hybrid production. *PLoS ONE*, 8, e61989. doi: 10.1371/journal.pone.0061989
28. Seren U, Vilhjálmsson BJ, Horton MW, Meng D, Forai P, Huang YS, Long Q, **Segura V**, Nordborg M. GWAPP : A Web Application for Genome wide Association Mapping in *A. thaliana*. *Plant Cell*, 24, 4793-4805. doi: 10.1105/tpc.112.108068

2012

29. Korte A*, Vilhjálmsson BJ*, **Segura V***, Platt A., Long Q, Nordborg M. A mixed-model approach for genome-wide association studies of correlated traits in structured populations. *Nature Genetics*, 44 : 1066-1071. doi: 10.1038/ng.2376
30. **Segura V***, Vilhjálmsson BJ*, Platt A, Korte A, Seren U, Long Q, Nordborg M. An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nature Genetics* 44 : 825-830. doi: 10.1038/ng.2314

*. contribution égale

2010

31. **Segura V**. Génétique et amélioration d'*Artemisia annua* L. pour une production durable d'antipaludiques à base d'artémisinine. *Médecine Sciences* 26, 701-703. doi: 10.1051/medsci/2010268-9701
32. Graham IA, Besser K, Blumer S, Branigan C, Czechowski T, Elias L, Guterma I, Harvey D, Isaac PG, Khan AM, Larson TR, Li Y, Pawson T, Penfield T, Rae AM, Rathbone DA, Reid S, Ross J, Smallwood MF, **Segura V**, Townsend T, Vyas D, Winzer T, Bowles DJ. The genetic map of *Artemisia annua* L. identifies loci affecting yield of the antimalarial drug artemisinin. *Science* 327, 328-331. doi: 10.1126/science.1182612
33. Fernandez L, Torregrosa L, **Segura V**, Bouquet A, Martinez-Zapater JM. Transposon induced gene activation as a mechanism generating cluster shape somatic variation in grapevine. *Plant Journal* 61, 545-557. doi: 10.1111/j.1365-313X.2009.04090.x

2009

34. **Segura V**, Durel CE, Costes E. Dissecting apple tree architecture into genetic, ontogenetic and environmental effects : QTL mapping. *Tree Genetics & Genomes* 5, 165-179. doi: 10.1007/s11295-008-0181-x

2008

35. **Segura V**, Cilas C, Costes E. Dissecting apple tree architecture into genetic, ontogenetic and environmental effects : mixed linear modelling of repeated spatial and temporal measures. *New Phytologist* 178, 302-314. doi: 10.1111/j.1469-8137.2007.02374.x
36. Bylesjö M, **Segura V**, Soolanayakanahally RY, Rae AM, Trygg J, Gustafsson P, Jansson S, Street NR. LAMINA : a tool for rapid quantification of leaf size and shape parameters. *BMC Plant Biology* 8, 82. doi: 10.1186/1471-2229-8-82
37. **Segura V**, Ouangraoua A, Ferraro P, Costes E. Comparison of tree architecture using tree edit distances : application to two-year-old apple hybrids. *Euphytica* 161, 155-164. doi: 10.1007/s10681-007-9430-6

2007

38. **Segura V**, Denancé C, Durel CE, Costes E. Wide range QTL analysis for complex architectural traits in a one-year-old apple progeny. *Genome* 50, 159-171. *New Phytologist* 178, 302-314. doi: 10.1139/G07-002

2006

39. **Segura V**, Cilas C, Laurens F, Costes E. Phenotyping progenies for complex architectural traits : a strategy for 1-year-old apple trees (*Malus x domestica* Borkh.). *Tree Genetics & Genomes* 2, 140-151. doi: 10.1007/s11295-006-0037-1

40. Legave JM, **Segura V**, Fournier D, Costes E. The effect of genotype, location and their interaction on early growth and branching in apricot trees. *Journal of Horticultural Science & Biotechnology* 81, 189-198. doi: 10.1080/14620316.2006.11512049

Chapitre 2

Rapport sur les travaux de recherche¹

Ce rapport présente une synthèse de mes travaux de recherche depuis ma thèse de doctorat jusqu'à aujourd'hui. Ces travaux portent sur l'architecture et le déterminisme génétique de caractères complexes chez différentes espèces, notamment pérennes, ainsi que sur des développements méthodologiques qui avaient pour objectifs de faciliter ces différentes études et de fournir des outils pour l'amélioration génétique.

2.1 Étude des déterminismes génétiques de caractères complexes par des approches de détection de QTLs

2.1.1 Déterminismes génétiques de l'architecture aérienne chez le pommier

Cette sous-partie concerne mon travail de thèse de doctorat au sein de l'équipe AFEF de l'UMR AGAP Institut (anciennement DAP au moment où s'est déroulé ce travail) à Montpellier.

Le contrôle du développement et de la forme des plantes est un enjeu important pour les espèces cultivées. Chez les arbres fruitiers, cet objectif est généralement atteint par l'utilisation de porte-greffes nanisants et la conduite culturale. Toutefois, ces pratiques peuvent présenter des coûts tels que ceux engendrés par les opérations de taille par exemple. L'introduction de caractères de forme et développement de l'arbre dans les schémas de sélection pouvait donc constituer une voie prometteuse afin de réduire les coûts liés à ces pratiques. Dans ce contexte, l'objectif de ma thèse était d'analyser les déterminismes génétiques des caractères architecturaux chez le pommier par une démarche de génétique quantitative. Cette approche était assez originale puisque peu de travaux de génétique avaient été rapportés sur cette thématique chez le pommier en particulier, mais aussi plus largement chez les arbres fruitiers (Conner et al., 1998; De Wit et al., 2004).

1. Les citations qui correspondent à des publications dont je suis co-auteur sont indiquées en gras dans le texte.

Une analyse prenant en compte les principaux processus architecturaux, mais aussi le développement pluri-annuel des arbres et la variabilité intra-arbre, avait été entreprise sur les 4 premières années de croissance d'une descendance F1 issue du croisement de deux variétés de pommiers. Le pommier étant une espèce hétérozygote, cette descendance présentait un certain niveau de variabilité phénotypique pour les caractères architecturaux. Des valeurs d'héritabilité au sens large moyennes à élevées ont été obtenues pour l'ensemble des caractères étudiés et un certain nombre de régions génomiques associées à ces caractères (QTL, Quantitative Trait Loci) ont été détectées par cartographie génétique (Segura et al., 2006, 2007). Cependant, en raison d'effets ontogéniques et climatiques, des variations ont été observées pour le nombre et les effets des QTLs identifiés. Afin de caractériser ces effets, les données ont été regroupées sous forme de séquences de mesures répétées dans le temps et l'espace, et analysées au moyen de modèles linéaires mixtes permettant de décomposer la plasticité phénotypique en effets génétiques, ontogéniques et environnementaux (Segura et al., 2008a). Des QTLs ont aussi pu être cartographiés pour chacun de ces effets, en travaillant directement à partir de leurs valeurs génétiques prédites (BLUP, Best Linear Unbiased Predictor) (Segura et al., 2009).

Par ailleurs, dans l'objectif de contribuer à la définition de critères pour les schémas d'amélioration génétique, des classifications ont été entreprises à la fois sur la base des descripteurs pertinents, mais aussi à partir de méthodes avancées de comparaison d'arborescences (Segura et al., 2006, 2008b).

À l'issue de ce travail, j'ai pu proposer une méthode de phénotypage et d'analyse, et cette démarche ouvrait des perspectives à la fois pour l'identification des déterminismes moléculaires sous-jacents aux QTLs mis en évidence, mais aussi pour la création variétale.

2.1.2 Génétique et amélioration d'*Artemisia annua* L. pour une production durable d'antipaludiques à base d'artémisinine

Cette sous-partie concerne mon travail post-doctoral au sein du projet Artemisia à l'Université de York au Royaume-Uni.

Le paludisme est un fléau majeur qui tue plus d'un million de personnes par an dans le monde. Le parasite responsable de cette maladie, *Plasmodium falciparum*, ayant développé des résistances envers la plupart des médicaments traditionnellement utilisés, comme la chloroquine, les associations médicamenteuses comportant de l'artémisinine (ACT, Artemisinin Combination Therapies) sont considérées comme les seuls traitements efficaces par l'OMS (Organisation Mondiale de la Santé). L'artémisinine est naturellement produite par l'espèce *Artemisia annua* L. de la famille des Astéracées. Il s'agit d'un sesquiterpène synthétisé au sein de groupes de cellules spécialisées, les trichomes glandulaires, situés principalement sur les feuilles. Cependant, la molécule demeure assez coûteuse car les plantes utilisées pour la production d'artémisinine présentent de très faibles rendements étant donné qu'elles n'avaient pas vraiment été sélectionnées dans un objectif de produc-

tion. Dans ce contexte, le projet Artemisia, dans lequel j'ai effectué mon contrat post-doctoral, ambitionnait accélérer la domestication d'*Artémisia annua* pour une production durable d'artémisinine.

Je travaillais au sein de l'équipe génétique et amélioration des plantes et mes missions concernaient principalement l'analyse de données collectées sur une descendance F1 correspondant à une des très rares variétés population disponibles sur cette espèce ("Artemis"). Comme dans le cas du pommier en effet, l'espèce *Artemisia annua* se reproduit majoritairement par fécondation croisée, ainsi les individus qui la composent sont généralement très hétérozygotes et les populations issues de leurs croisements présentent une variabilité en F1 exploitable pour des analyses génétiques. Plus spécifiquement, j'ai contribué à la définition d'une stratégie de phénotypage intégrant analyses du développement des plantes et de la concentration en métabolites dans les feuilles pour décomposer le rendement global en artémisinine. J'ai participé aux collectes de données phénotypiques et je les ai analysées au moyen de modèles linéaires mixtes pour prendre notamment en compte des effets environnementaux et leurs interactions avec le génotype puisque les données avaient été collectées sur plusieurs sites sur du matériel cloné par bouturage. J'ai ensuite entrepris des détections de QTLs à partir des valeurs génotypiques prédites et j'ai notamment pu détecter des QTL stables en fonction de l'environnement pour la masse fraîche des plantes, la surface foliaire, l'architecture des plantes et la concentration en artémisinine dans les feuilles. De façon notable, j'ai par ailleurs pu valider un de ces QTLs dans une population mutante, obtenue par traitement chimique de la variété "Artemis". En effet, l'analyse de la ségrégation de marqueurs moléculaires chez des individus mutants sélectionnés pour leur rendement en artémisinine a révélé une très forte distorsion en faveur de l'allèle agissant positivement sur la concentration en artémisinine dans la région génomique de ce QTL. Ces résultats ont été inclus dans un article plus large correspondant à l'ensemble des avancées du projet depuis le développement de marqueurs moléculaires et l'étude de l'expression de gènes par séquençage de banques d'ADN complémentaire (ADNc), jusqu'à la détection de QTLs (Graham et al., 2010).

A l'issue de ce travail le projet disposait d'un certain nombre d'outils (phénotypage, marqueurs moléculaires) pour accélérer la domestication de l'espèce. Ces travaux devaient se poursuivre par de la génétique d'association dans des populations naturelles, mais cela a fait l'objet du travail d'autres collègues puisque j'ai dû quitter le projet suite à la réussite du concours chargé de recherche INRA pour un profil intitulé "Biologie intégrative de la production de biomasse chez le peuplier" à l'UMR BioForA (anciennement AGPF) à Orléans.

2.1.3 Bilan

Ces deux expériences de recherche m'ont permis d'acquérir une expertise dans l'étude de la variabilité et du déterminisme génétique de caractères complexes principalement au sein de croisements biparentaux entre espèces hétérozygotes, depuis la définition et la mise en œuvre du phénotypage jusqu'à l'analyse statistique des don-

nées et leur interprétation et valorisation. J'ai notamment pu observé une certaine tendance concernant le succès des approches de détection de QTLs : la décomposition de phénotypes complexes en phénotypes plus proches des processus biologiques sous-jacents semble être un facteur important pour le succès de ces approches, sans pour autant que cela en soit une garantie. Cela peut s'expliquer par le fait que des caractères complexes, tels que le port de l'arbre ou le rendement global en une molécule, résultent de la combinaison de plusieurs phénomènes biologiques eux-mêmes sous déterminisme potentiellement complexe et polygénique.

2.2 Développements méthodologiques pour la GWAS et collaborations associées

Suite à mon recrutement en tant que chargé de recherche, j'ai eu l'opportunité d'effectuer une mission longue durée à l'étranger d'un an dans l'équipe de Magnus Nordborg au Gregor Mendel Institute for Molecular Plant Biology (GMI) à Vienne en Autriche. Cette équipe était spécialisée dans les études d'association pangénomiques (GWAS - Genome Wide Association Studies) chez les plantes (Zhao et al., 2007; Atwell et al., 2010; Platt et al., 2010) et l'objectif initial de ma mission était donc de me former à ce type d'approches afin d'acquérir une expertise que je pourrai ensuite déployer dans le cadre de mes projets de recherche. Au-delà de cette initiation aux GWAS, j'ai activement participé à deux développements méthodologiques originaux qui correspondent à des extensions du modèle de détection d'association, pour la prise en compte simultanée de plusieurs loci ou l'analyse conjointe de plusieurs caractères ou environnements.

2.2.1 Le modèle linéaire mixte pour la génétique d'association

Ces méthodes sont basées sur le modèle polygénique ou modèle animal qui est de nos jours encore considéré comme le modèle par défaut en génétique d'association. Ce modèle a initialement été proposé par Fisher au début du 20^{ème} siècle (Fisher, 1919), et il a été formalisé en modèle mixte par Henderson dans les années 1970 pour devenir un outil indispensable aux sélectionneurs animaliers (Henderson, 1984). L'utilisation de ce modèle en génétique d'association a été initialement proposé par Yu et al. (2006) pour prendre en compte l'effet confondant généralement attribué à la structure des populations et à l'apparentement entre individus. En effet, dans le cas relativement fréquent où la population au sein de laquelle l'analyse d'association entre polymorphismes et phénotype est structurée en sous-groupes génétiques, l'application d'un modèle naïf de régression linéaire simple conduit à la détection d'un très (trop) grand nombre de polymorphismes significatifs même après correction des p-valeurs pour le fait d'avoir effectué un grand nombre de tests. Ce phénomène se produit notamment lorsque le phénotype d'intérêt est lui même structuré, c'est à dire qu'il présente une certaine différenciation entre les sous-populations. Dans ce cas de figure, tout polymorphisme associé à cette différenciation génétique sera as-

socié au phénotype sans pour autant qu'il soit génétiquement lié au polymorphisme causal. Si ce phénomène avait bien été identifié et documenté en génétique humaine, comme en témoigne cette revue de littérature par Cardon and Palmer (2003), les approches proposées par cette communauté pour y remédier comme le contrôle génomique (Devlin and Roeder, 1999) ou l'association stratifiée (Pritchard et al., 2000) n'étaient pas tout à fait satisfaisantes et le modèle linéaire mixte pour la génétique d'association proposé en génétique végétale a finalement aussi fini par être adopté en génétique humaine quelques années plus tard (Price et al., 2010; Yang et al., 2014).

Ce modèle peut s'écrire de la façon suivante :

$$\mathbf{y} = \beta_0 + \mathbf{Z}\mathbf{u} + \varepsilon \quad (2.1)$$

- \mathbf{y} est un vecteur de phénotypes ;
- β_0 est l'intercept qui permet d'ajuster la moyenne générale du phénotype ;
- \mathbf{Z} est une matrice d'incidence reliant les observations aux individus, il s'agit d'une matrice d'identité lorsque les phénotypes sont considérés à l'échelle génotypique ;
- \mathbf{u} est un vecteur d'effets génétiques aléatoires, $\mathbf{u} \sim \mathcal{N}(0, \sigma_u^2 \mathbf{K})$, et \mathbf{K} est une matrice de covariance entre génotypes (voir ci-après) ;
- ε est un vecteur de résidus, $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2 \mathbf{I})$.

On peut noter que ce modèle linéaire mixte (2.1) présente une équivalence avec le modèle de régression aléatoire dans lequel tous les polymorphismes sont inclus simultanément dans le modèle :

$$\mathbf{y} = \beta_0 + \mathbf{X}\beta + \varepsilon \quad (2.2)$$

- \mathbf{X} est la matrice de polymorphismes (typiquement des SNPs, Single Nucleotide Polymorphisms) codés numériquement selon l'hypothèse d'additivité d'un allèle par rapport à l'autre : $\forall i, p x_{ip} \in \{0, 1, 2\}$ correspondant au nombre de doses alléliques de l'allèle alternatif par rapport à l'allèle de référence pour le SNP p chez l'individu i ;
- β est un vecteur d'effets aléatoires, $\beta \sim \mathcal{N}(0, \mathbf{I}\sigma_\beta^2)$.

Ce modèle est généralement résolu par une approche de régression pénalisée (ridge regression) puisque il y a habituellement plus de SNPs à tester que d'individus évalués ($p > i$). Dans ce cas pour qu'il y ait équivalence entre les deux modèles, il faut que le paramètre de régularisation (λ) dans le modèle de régression pénalisée soit égal au ratio $\sigma_\varepsilon^2/\sigma_u^2$ du modèle 2.1. Il faut aussi que la matrice de covariance \mathbf{K} dans le modèle 2.1 soit estimée à partir de la matrice de SNPs de la façon suivante :

$$\mathbf{K} = \frac{\mathbf{X}\mathbf{X}^T}{2 \sum_p f_p(1 - f_p)} \quad (2.3)$$

où f_p est la fréquence de l'allèle alternatif pour le SNP p . L'équivalence entre les deux modèles se retrouve alors au niveau des variances : $\sigma_u^2 = 2 \sum_p f_p(1 - f_p)\sigma_\beta^2$, c'est à

dire que la variance génétique dans le modèle 2.1 se retrouve partagée entre tous les SNPs dans le modèle 2.2. Cela permet de comprendre pourquoi l'effet génétique \mathbf{u} est souvent qualifié d'effet polygénique.

Si on revient à la génétique d'association, Yu et al. (2006) ont proposé le modèle suivant :

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \varepsilon$$

$$\mathbf{y} = \underbrace{\beta_0 + \mathbf{x}_p\alpha_p + \mathbf{Q}\mathbf{v}} + \mathbf{Z}\mathbf{u} + \varepsilon \quad (2.4)$$

où \mathbf{x}_p et \mathbf{Q} sont les matrices d'incidences reliant les individus aux effets fixes du marqueur p à tester, α_p , et de la structure génétique, \mathbf{v} . On peut noter que dans de nombreuses applications (le plus souvent), il n'est pas indispensable d'inclure d'effet fixe de la structure génétique dans le modèle, puisque cet effet est généralement capté par l'effet génétique aléatoire ou effet polygénique. D'ailleurs l'identification de sous-populations et l'assignation des génotypes à chacune de ces sous-populations se fait habituellement à partir des SNPs, et de fait les deux parties du modèle, fixes et aléatoires, présentent une certaine redondance d'information qui n'est généralement pas une situation favorable pour l'ajustement du modèle (Astle and Balding, 2009). Parfois même, certaines études utilisent comme matrice d'incidence (\mathbf{Q}) les coordonnées des individus sur les axes d'une ACP (Analyse en Composantes Principales) effectuée à partir des données de SNPs (Price et al., 2006). Or généralement cette ACP correspond à une décomposition de la matrice de covariance \mathbf{K} entre les génotypes estimées à partir des SNPs. Dans ce cas précis, la redondance d'information est flagrante notamment dans le cas où le nombre d'axes d'ACP retenu pour la partie fixe est élevé et de fait cette stratégie d'analyse semble sous-optimale. Néanmoins, l'observation empirique de la distribution des p-valeurs, au moyen par exemple d'un qq-plot (quantile-quantile plot) reste à mon avis le meilleur moyen de diagnostiquer l'effet confondant de la structure et si après application d'un modèle avec seulement l'effet génétique aléatoire les p-valeurs ne suivent pas majoritairement une loi uniforme, il est conseillé de tester un modèle avec une prise en compte explicite de la structure en effet fixe. Cette situation est notamment rencontrée dans les cas où la structure de la population étudiée est forte (par exemple lorsque l'analyse est effectuée sur des sous-espèces) et surtout lorsque les caractères étudiés sont particulièrement différenciés entre les sous-populations.

En pratique la détection d'associations se fait en répétant le modèle 2.4 sur tous les SNPs disponibles et en testant à chaque fois l'effet du SNP p : $H_1 : \alpha_p \neq 0$ contre $H_0 : \alpha_p = 0$. Cela conduit à effectuer autant de tests qu'il y a de SNPs et il convient de le prendre en compte au moyen d'une approche dédiée comme la correction de Bonferroni par exemple (Balding, 2006).

2.2.2 MLMM

Si le modèle mixte fonctionne bien pour contrôler l'effet confondant de la structure des populations, l'inclusion de plusieurs SNPs comme cofacteurs dans le modèle apparaissait comme une évidence pour plusieurs raisons :

- lorsque dans l'architecture génétique du caractère il y a des loci à effet relativement fort, ils peuvent générer un effet confondant qui est mal pris en compte par l'effet polygénique, puisque ce dernier fait l'hypothèse que les effets sont distribués selon une loi normale ;
- cette pratique est commune dans les études de cartographie de QTLs via les approches de cartographie de QTLs multiples (MQM - multiple qtl mapping) (Jansen, 1993) ou de cartographie par intervalles composites (CIM - composite interval mapping) (Zeng, 1994) dont les avantages par rapport à l'approche de cartographie par intervalle simple (SIM - simple interval mapping) ont été démontrés ;
- en génétique d'association pangénomique, quelques études avaient montré que le fait de mettre en cofacteur le SNP le plus significatif pouvait permettre de mieux appréhender certains pics d'association en identifiant notamment des situations d'hétérogénéité allélique (Lango Allen et al., 2010) ;
- une étude basée sur des simulations avait montré que l'utilisation de SNPs candidats comme cofacteurs dans le modèle de détection permettait d'accroître le puissance du modèle (Ma et al., 2010).

Sur la base de ces constats, nous avons donc proposé d'étendre le modèle linéaire mixte utilisé en génétique d'association pour prendre en compte plusieurs loci comme cofacteurs en combinant modèle mixte approximatif (Kang et al., 2010) (car les approches exactes n'étaient pas suffisamment rapides à ce moment là pour analyser les gros jeux de données classiquement utilisés en GWAS) et régression multiple pas à pas. Ce nouveau modèle a été appelé MLM pour Multi-Locus Mixed-Model (**Segura et al., 2012**). En pratique, il s'agit d'effectuer une "inclusion forward" pour un nombre d'itérations pré-déterminées puis de réaliser une "exclusion backward". A chaque itération, les composantes de variance (σ_u^2 et σ_ε^2) sont ré-estimées ce qui minimise l'impact d'avoir utilisé un modèle mixte approximatif pour effectuer les scans génomiques. A fur et à mesure que des SNPs sont inclus comme cofacteurs dans la partie fixe du modèle, la variance attribuée au terme polygénique tend à décroître de la partie effectivement capturée par les SNPs cofacteurs. L'inclusion forward est donc automatiquement stoppée lorsque l'héritabilité génomique ($h^2 = \sigma_u^2 / (\sigma_u^2 + \sigma_\varepsilon^2)$) vaut 0, même si le nombre d'itérations pré-déterminées n'a pas été atteint. A l'issue de ces itérations (inclusion forward et élimination backward), un certain nombre de modèles ont été ajustés et il convient de choisir parmi ces modèles celui qui est le plus pertinent. Pour cela nous avons proposé d'utiliser deux critères : le BIC (critère d'information bayésien) étendu (eBIC) qui correspond à un BIC sur-pénalisé par l'étendue des possibles (Chen and Chen, 2008) et un critère de Bonferroni multiple (mBonf) qui cherche le modèle avec le plus de cofacteurs tous significatifs après correction de Bonferroni.

Pour évaluer la pertinence de MLM, nous avons tout d'abord effectué des simulations à partir de données génotypiques réelles d'*Arabidopsis thaliana* (Horton et al., 2012), puis nous l'avons appliqué à des jeux de données réels précédemment analysés avec le modèle mixte simple locus notamment. Les simulations ont montré

que MLMM surpasse notamment le modèle mixte simple locus à la fois en terme de pouvoir de détection mais aussi de fausses découvertes (FIGURE 2). Cela est d'autant plus le cas que l'architecture du caractère étudié présente des loci à effet relativement fort. Elles montrent aussi que les critères proposés pour sélectionner le meilleur modèle semblent bien pertinents, le critère eBIC ayant tendance à sélectionner moins de SNPs cofacteurs que le critère mBonf. Par ailleurs, l'application du modèle à des données réelles d'*Arabidopsis thaliana* ou de l'Homme a permis d'identifier de nouvelles associations, dont notamment des cas d'hétérogénéité allélique, c'est-à-dire des associations multiples au sein d'un même locus, ce qui souligne son intérêt.

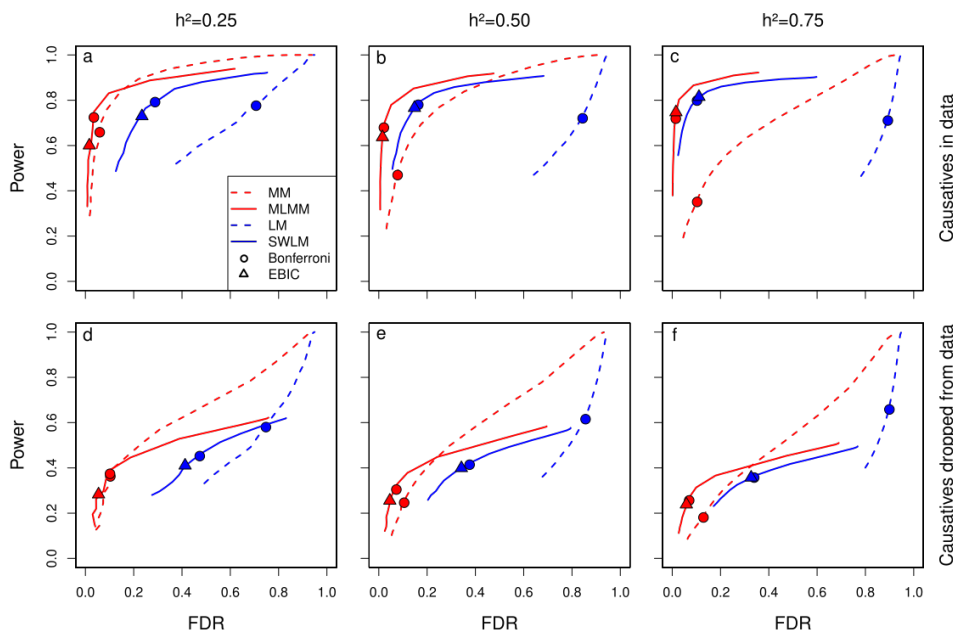


FIGURE 2 – Résultats des simulations effectuées pour évaluer les performances de la méthode MLMM.

Performances respectives de quatre méthodes de cartographie sur données simulées : régression linéaire simple (LM), régression multiple pas à pas (SWLM), modèle linéaire mixte simple locus (MM) et modèle linéaire mixte multi-locus (MLMM). Les cercles et les triangles représentent les modèles identifiés avec les critères mBonf et EBIC, respectivement. Puissance et FDR ont été estimés avec (a) et sans (b) les loci causaux inclus dans le jeu de données. Trois héritabilités phénotypiques ont été utilisées dans les simulations : 0.25 (gauche), 0.5 (milieu) et 0.75 (droite).

L'application de ce modèle en génétique végétale s'est depuis bien popularisée notamment grâce à la mise à disposition de la communauté de codes Python et R. On peut ici mentionner la contribution de collègues de l'INRA (notamment Timothée Flutre et Brigitte Mangin) pour mettre le code R sous forme de package ce qui a grandement facilité la diffusion de l'approche et a vraisemblablement contribué de façon significative à sa popularité (Bonnafous et al., 2019).

Aujourd'hui avec près de 10 ans de recul sur cette approche, je peux confirmer que la méthode MLMM est "loin d'être une panacée", pour reprendre les termes

utilisés dans la discussion de l'article original (Segura et al., 2012). En effet, l'approche semble surtout très utile pour mieux comprendre l'architecture génétique des caractères et prendre en compte les loci à effet relativement fort. C'est dans ce type de situation qu'elle apporte une réelle plus-value, mais elle ne va pas faire de "miracle" s'il n'y a pas vraiment de signal en simple locus dans le jeu de données. D'un point de vue plus statistique, l'approche est somme toute assez simple et c'est vraisemblablement ce qui l'a rendue si populaire auprès des généticiens végétaux. En effet, depuis un certain nombre d'autres approches potentiellement plus puissantes ont été rapportées dans la littérature (Liu et al., 2016; Huang et al., 2019), mais il se pourrait que la complexité des modèles sous-jacents ait tendance à les rendre un peu moins accessibles à la communauté qui généralement aime bien comprendre ce qu'elle met en œuvre.

2.2.3 MTMM

Pour ce qui est de l'approche multi-caractères, la motivation venait surtout de la volonté d'effectuer de la génétique d'association dans un contexte multi-environnemental, dans lequel le même caractère mesuré dans plusieurs conditions peut-être considéré comme plusieurs caractères mesurés dans une même condition. Dans ce cas particulier, l'intérêt de cette approche consiste surtout à tester spécifiquement un effet d'interaction entre environnement et SNPs qui peut permettre de détecter des loci potentiellement adaptatifs car impliqués dans la réponse des plantes à l'environnement et donc potentiellement explicatifs de l'IGE (Interaction entre Génotype et Environnement). Une situation assez favorable pour appliquer ce type d'approche correspond au cas où un facteur particulier est appliqué à la population d'étude en vue d'étudier spécifiquement la réponse des génotypes à ce facteur, qui peut être par exemple une contrainte hydrique ou un traitement chimique.

Le modèle proposé, appelé MTMM (Multi-Trait Mixed-Model, (Korte et al., 2012)), est simplement une extension multivariée du modèle linéaire mixte pour la génétique d'association (Yu et al., 2006). Dans le cas par exemple de 2 caractères ou 2 environnements, le modèle peut s'écrire de la façon suivante :

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \mathbf{s}_1\mu_1 + \mathbf{s}_2\mu_2 + \mathbf{x}_p\alpha_p + (\mathbf{x}_p \times \mathbf{s}_1)\gamma_p + \mathbf{Z}\mathbf{u} + \varepsilon \quad (2.5)$$

- $\mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix}$ est le vecteur de phénotypes ;
- \mathbf{s}_j sont des matrices d'incidence reliant les observations aux caractères ou environnements j , et μ_j sont les effets fixes correspondant ;
- \mathbf{x}_p et $(\mathbf{x}_p \times \mathbf{s}_1)$ sont respectivement les matrices d'incidences reliant les observations à l'effet global du SNP α_p et à son effet d'interaction avec le caractère ou l'environnement γ_p ;

- \mathbf{Z} est une matrice d'incidence reliant les observations aux individus et \mathbf{u} est l'effet aléatoire génétique, défini par $\mathbf{u} \sim \mathcal{N}(0, \begin{bmatrix} \sigma_{u_1}^2 & \sigma_{u_{12}} \\ \sigma_{u_{12}} & \sigma_{u_2}^2 \end{bmatrix} \otimes \mathbf{K})$;
- ε est la résiduelle, $\varepsilon \sim \mathcal{N}(0, \begin{bmatrix} \sigma_{\varepsilon_1}^2 & \sigma_{\varepsilon_{12}} \\ \sigma_{\varepsilon_{12}} & \sigma_{\varepsilon_2}^2 \end{bmatrix} \otimes \mathbf{I})$.

Les paramètres de variance et covariance des effets aléatoires permettent d'estimer les héritabilité génomiques (h_j^2) des caractères étudiés ainsi que leurs corrélations génétiques (ρ_u) et résiduelles (ρ_ε).

$$\begin{aligned} h_j^2 &= \sigma_{u_j}^2 / (\sigma_{u_j}^2 + \sigma_{\varepsilon_j}^2) \\ \rho_u &= \sigma_{u_{12}} / \sqrt{\sigma_{u_1}^2 \sigma_{u_2}^2} \\ \rho_\varepsilon &= \sigma_{\varepsilon_{12}} / \sqrt{\sigma_{\varepsilon_1}^2 \sigma_{\varepsilon_2}^2} \end{aligned} \quad (2.6)$$

On peut d'ailleurs noter que dans le cas de figure où on s'intéresse à des caractères mesurés sur les mêmes génotypes dans des conditions environnementales différentes, la covariance résiduelle peut-être considérée comme nulle, ce qui diminue le nombre de paramètres à estimer. Par ailleurs, dans cette même situation, on peut également déduire de ces différentes estimations les variances génétique et d'IGE (Itoh and Yamada, 1990).

Pour permettre la faisabilité du modèle MTMM dans un contexte pangénomique, l'approche modèle-mixte approximatif, proposée en univarié par Kang et al. (2010), a été adoptée pour effectuer les tests entre les modèles imbriqués suivants :

- Modèle complet : $\mathbf{s}_1\mu_1 + \mathbf{s}_2\mu_2 + \mathbf{x}_p\alpha_p + (\mathbf{x}_p \times \mathbf{s}_1)\gamma_p + \mathbf{Z}\mathbf{u} + \varepsilon$
- Modèle réduit : $\mathbf{s}_1\mu_1 + \mathbf{s}_2\mu_2 + \mathbf{x}_p\alpha_p + \mathbf{Z}\mathbf{u} + \varepsilon$
- Modèle null : $\mathbf{y} = \mathbf{s}_1\mu_1 + \mathbf{s}_2\mu_2 + \mathbf{Z}\mathbf{u} + \varepsilon$

Le test entre modèle complet et modèle null renseigne de l'effet global du SNP sur les 2 caractères ou environnements, tandis que le test entre modèle complet et modèle réduit renseigne de l'effet d'interaction entre SNP et caractère ou SNP et environnement. Enfin un test entre modèle complet et modèle null, rassemble l'information des deux tests précédents et ainsi informe de l'existence d'un effet quelle que soit sa nature.

Comme dans le cas de MLM, les performances de MTMM ont d'abord été étudiées par simulations à partir de données génotypiques réelles d'*Arabidopsis thaliana* (Horton et al., 2012). Ces simulations soulignent clairement l'intérêt du modèles (FIGURE 3) notamment en présence de corrélations génétiques entre les caractères étudiés qui suggérerait l'existence de loci sous-jacents partagés (situations de pléiotropie ou "linkage"). Dans un deuxième temps, nous avons appliqué MTMM à des jeux de données réelles d'*A. thaliana* ou de l'Homme. Nous avons notamment pu identifier des associations qui n'avaient pas été détectées précédemment mais qui semblaient bien pertinentes, soit du point de vue de l'annotation des gènes situés à proximité, soit parce que des méta-analyses incluant de très larges ensembles de

cohortes les avaient par ailleurs également détectées.

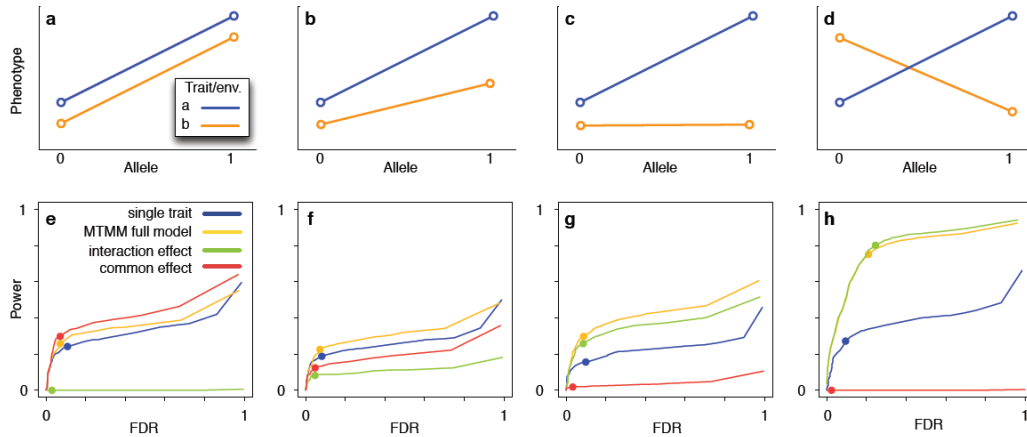


FIGURE 3 – Résultats des simulations effectuées pour évaluer les performances de la méthode MTMM.

(a–d) Scénarios simulés, avec pléiotropie positive ou effet commun à plusieurs environnements (a); pléiotropie positive ou effet commun à plusieurs environnements, avec taille d'effet différente selon le caractère ou l'environnement (b); effet seulement pour un caractère ou environnement (c); pléiotropie négative ou effet opposé selon l'environnement (d). (e–f) Performances de six méthodes différentes dans les scénarios précédemment décrits. Le point sur les courbes dans les panels e à h correspond au seuil de significativité de 5% après correction selon la méthode de Bonferroni.

Malgré tout son intérêt, cette approche semble un peu moins populaire que MLM dans la littérature et plusieurs hypothèses peuvent être avancées pour expliquer cela :

- l'implémentation originale reposait sur l'utilisation d'un logiciel payant ASReML-R (Butler et al., 2018), car au moment de son développement il n'y avait pas beaucoup d'alternatives libres ;
- en situation multi-caractères (mesurés sur un même individu), d'autres alternatives efficaces ont été développées depuis en génétique humaine comme par exemple le logiciel Gemma (Zhou and Stephens, 2014) ;
- l'application à des situations multi-environnementales pour étudier la réponse à une contrainte en génétique végétale est tout de même bien moins répandue que l'approche de GWAS classique simple-environnement.

Pour pallier notamment à la limite de MTMM mentionnée ci-dessus concernant la dépendance au logiciel payant ASReML-R, d'autres alternatives existent désormais pour ajuster le modèle mixte multivarié décrit ci-dessus (2.5), comme par exemple les packages `breedR` (Muñoz and Sanchez, 2021) et `sommer` (Covarrubias-Pazarán, 2018). Ces alternatives ouvrent des perspectives pour une application plus systématique du modèle MTMM dans les situations où l'on s'intéresse spécifiquement à identifier des loci impliqués dans la réponse des génotypes à un facteur particulier.

2.2.4 Collaborations associées

Les avancées technologiques en matière de séquençage combinées notamment aux stratégies de réduction de complexité des génomes par enzymes de restriction ou capture (Elshire et al., 2011) ont permis de démocratiser le génotypage à haut-débit chez de nombreuses espèces ouvrant des perspectives de détection de QTL par génétique d'association. Mon expertise en matière de méthodologie pour la génétique d'association m'a ainsi offert de nombreuses collaborations principalement au sein de l'INRA, dont certaines ont notamment abouti à des valorisation dont je suis co-auteur sur tomate (Sauvage et al., 2014; Albert et al., 2016; Bauchet et al., 2017), maïs (Giraud et al., 2014), vigne (Fodor et al., 2014), pommier (Coupel-Ledru et al., soumis), rouille du peuplier (Persoons et al., soumis).

Parmi ces collaborations, certaines ont nécessité des modifications par rapport aux modèles initiaux. C'est le cas de l'application sur maïs qui était basée sur un dispositif multi-parental plutôt que sur un panel de diversité et il a donc fallu adapter MLM pour ce type de dispositif (Giraud et al., 2014). Dans le cas de la tomate, il a fallu modifier le code pour permettre l'utilisation d'un effet fixe de la structure puisque le panel utilisé incluait plusieurs sous-espèces et de fait le terme polygénique n'était pas suffisant pour contrôler l'effet de la structure (Sauvage et al., 2014; Albert et al., 2016; Bauchet et al., 2017). Par ailleurs, le niveau de déséquilibre de liaison étant relativement important chez cette espèce, la correction de Bonferroni pour les tests multiples et son adaptation pour MLM (critère mBonf) ne semblaient pas optimales. J'ai donc modifié le code pour permettre à l'utilisateur de fournir son propre seuil de p-valeur pour l'identification du meilleur modèle.

2.3 Biologie intégrative de la production de biomasse chez le peuplier

Cette partie concerne le projet de recherche qui m'a été confié lors de mon recrutement en tant que chargé de recherche au sein de l'UMR BioForA (anciennement AGPF) à Orléans, et que j'ai réellement développé au retour de ma mission longue durée en Autriche. Ce projet a notamment bénéficié du soutien financier de l'agence nationale de la recherche via le financement du projet jeune chercheur SYBIOPOP (ANR-13-JSV6-0001) que j'ai coordonné.

2.3.1 Contexte

Le peuplier est une essence forestière d'importance écologique considérable puisqu'il fait partie des espèces dominantes des forêts alluviales (ripisylves) sous nos latitudes. Il présente par ailleurs une certaine importance économique puisqu'il est cultivé sur environ 200 000 ha de plantations monoclonales en France, la valorisation de son bois se faisant majoritairement sous forme d'emballage léger (cagettes, bouchons, barquettes...). Pour répondre aux enjeux posés par les contraintes biotiques

et abiotiques et contribuer au maintien d'une production conséquente de bois de qualité pour la filière, l'unité de recherche BioForA à Orléans a développé depuis de nombreuses années un programme d'amélioration génétique du Peuplier qui repose sur la valorisation de la variabilité génétique rassemblée au sein de 3 populations de base représentant les 3 espèces : *Populus deltoides*, *Populus trichocarpa* et *Populus nigra*. Les deux premières espèces sont d'origine nord-américaine, tandis que la dernière est d'origine eurasiennne. Ces espèces sont inter-fertiles et leurs hybrides, particulièrement vigoureux, sont la base de la populiculture traditionnelle.

Un intérêt grandissant existe par ailleurs pour la valorisation de la biomasse forestière sous forme d'énergie, et le peuplier de part sa croissance rapide constitue une espèce de choix dans cet objectif. Pour cela, un système de culture dédié a été proposé : le taillis à courte et à très courte rotation. Ce système original est radicalement différent de celui utilisé en populiculture classique. De ce fait, les variétés améliorées pour la production de bois d'œuvre ne sont pas nécessairement les plus appropriées pour la production de biomasse, malgré l'existence de critères d'amélioration communs tels que la productivité, la résistance à la rouille foliaire ou l'efficacité d'utilisation de l'eau et des éléments minéraux. En effet, jusqu'à présent les programmes de sélection n'ont pas pris en compte des critères visant à optimiser à la fois la quantité et la qualité de la production lignocellulosique, tels que l'aptitude au rejet de souche, la tolérance à la compétition et les propriétés chimiques du bois en vue de sa valorisation sous forme d'énergie comme par exemple le bioéthanol.

2.3.2 Questions de recherche et plan de travail

Mon travail sur peuplier était centré sur une collection de peupliers noirs (espèce *Populus nigra*) originaires de populations naturelles et représentative de l'aire de répartition de l'espèce en Europe de l'Ouest. L'objectif de mes recherches sur cette collection était d'étudier l'architecture génétique des caractères de production et de qualité du bois, en vue de transférer ces informations au programme d'amélioration génétique. Plus spécifiquement, je cherchais à répondre aux questions suivantes :

- Existe-t-il de la variabilité génétique pour les caractères cibles dans les populations naturelles de peuplier noir ?
- Comment est structurée cette variabilité génétique au sein des populations naturelles ?
- Est-ce que les caractères de qualité covarient avec les caractères de production, suggérant d'éventuels compromis sélectifs ?
- Comment l'environnement module-t-il cette variabilité en interaction avec le génotype ?
- Quels sont les déterminismes sous-jacents à la variabilité génétique observée pour les caractères de production et de qualité du bois ?

Pour répondre à ces questions de recherche, une démarche couplant phénotypage à haut-débit, génétique quantitative et transcriptomique avait été initiée dans le cadre du projet SYBIOPOP sur la collection de peuplier noir (FIGURE 4).

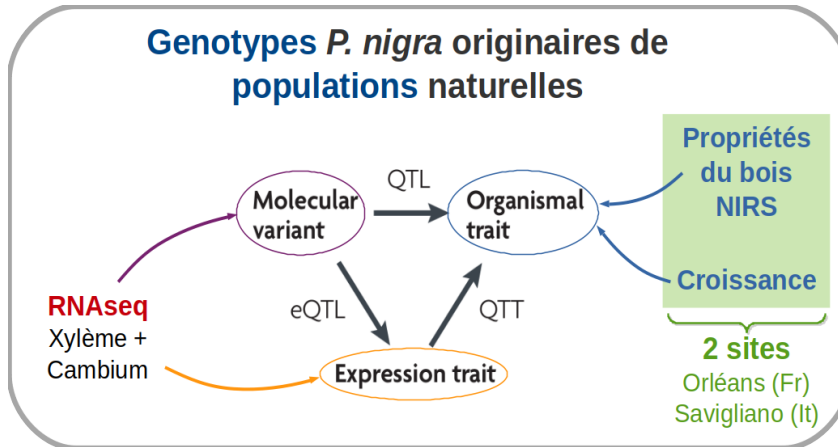


FIGURE 4 – Illustration de la démarche mise en œuvre dans le cadre du projet SYBIOPOP.

Le projet repose sur une collection de peupliers noirs évalués dans deux jardins communs (Orléans et Savigliano). Le phénotypage concerne des caractères de croissance et de propriété du bois prédits par spectrométrie dans le proche infrarouge (NIRS). Les données moléculaires, SNPs et transcrits, proviennent de séquençage d'ARNm (RNAseq) de xylème et cambium. L'objectif final étant de relier SNPs, transcrits et phénotypes, en identifiant QTL, QTL d'expression (eQTL) et transcrits associés aux caractères (QTT), pour mieux comprendre l'architecture génétique des caractères étudiés.

Un certain nombre de travaux avaient été rapportés sur les déterminismes génétiques de la croissance et des propriétés chimiques du bois chez le peuplier dans la littérature (Wegrzyn et al., 2010; Guerra et al., 2013; Porth et al., 2013; Allwright et al., 2016; Fahrenkrog et al., 2017), mais d'une façon générale les loci identifiés n'expliquaient qu'une faible part de la variabilité génétique des caractères (notion d'héritabilité manquante (Manolio et al., 2009)). Aussi pour résoudre spécifiquement ce problème, j'avais fait l'hypothèse que d'utiliser des données intermédiaires entre génome et phénotype, comme l'expression des gènes, devrait pouvoir permettre de décomplexifier l'architecture génétique des caractères d'intérêt afin de mieux comprendre et appréhender leurs déterminismes moléculaires.

La collection de peupliers noirs était composée de plus de 1 000 génotypes organisés en une dizaine d'origines géographiques ou populations en Europe de l'Ouest (FIGURE 5). Ces génotypes ont été clonés par bouturage afin de les implanter dans 2 dispositifs de type jardin commun, localisés respectivement à Orléans et Savigliano (Nord de l'Italie). Les dispositifs comprenaient 6 blocs complets randomisés, c'est à dire que dans chaque site les génotypes étaient répétés six fois. Au total, sur les 2 sites, on dénombrait près de 700 génotypes communs. Un certain nombre de caractères, ciblant les processus de croissance, phénologie, architecture de l'arbre,

avaient été évalués dans chacun des dispositifs dans le cadre de projets précédents. Par ailleurs, une puce de génotypage avait été développée et utilisée sur la collection fournissant environ 8 000 SNPs pour réaliser des études d'association avec les caractères (Faivre-Rampant et al., 2016).

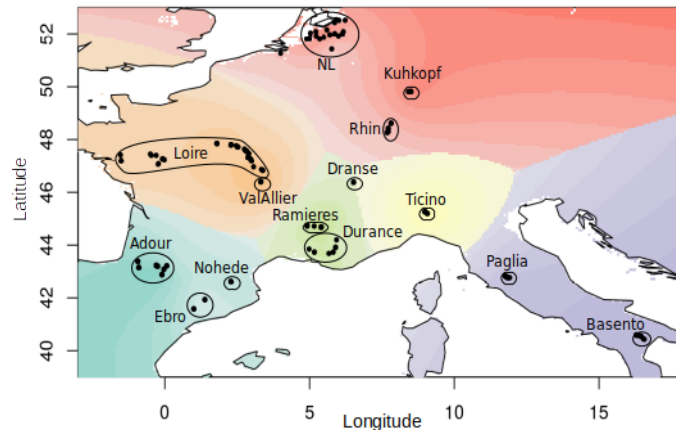


FIGURE 5 – Représentation graphique de l'origine géographique et de la structure de la diversité génétique de la collection de peupliers noirs.

Les points représentent les génotypes et les ellipses définissent les groupes géographiques ou populations, tandis que les couleurs représentent les populations ancestrales identifiées par analyse des données de génotypage SNP avec le logiciel admixture. Une seule population ancestrale n'apparaît pas sur le graphique, il s'agit de celle correspondant au compartiment cultivé (peuplier d'Italie).

Pour mener à bien le projet, j'ai mis en œuvre différentes actions afin de compléter ces jeux de données disponibles et d'étudier la variabilité génétique des caractères cibles :

- phénotypage à haut-débit et étude de la variabilité des propriétés du bois ;
- séquençage d'ARN (RNAseq) à partir de xylème et cambium collectés dans le dispositif orléanais avec comme double objectif de densifier le génotypage et de disposer du niveau d'expression de l'ensemble des gènes exprimés dans les tissus échantillonnés ;
- développement d'un pipeline de détection et de génotypage des SNPs à partir des données du RNAseq ;
- étude de la variabilité du transcriptome, construction de réseaux de co-expression de gènes et prédiction de caractères ;
- analyses de génétique d'association à l'échelle du transcriptome ;
- intégration de données génomiques et transcriptomiques pour la prédiction de phénotypes.

2.3.3 Principaux résultats

2.3.3.1 Variabilité génétique des propriétés du bois évaluées par spectrométrie dans le proche infrarouge

En ce qui concerne la variabilité des propriétés du bois, un très grand nombre d'échantillons ($\approx 6\ 000$) avaient été collectés dans les 2 dispositifs expérimentaux à l'occasion de coupes puisque les arbres étaient conduits en taillis à courte rotation (2 récoltes à Orléans et 1 récolte à Savigliano). Étant donné qu'il aurait été impossible d'évaluer tous ces échantillons dans des temps et pour des coûts raisonnables avec les méthodes biochimiques classiques, j'ai utilisé la spectroscopie dans le proche infrarouge (NIRS, Near InfraRed Spectroscopy) comme méthode de phénotypage à haut-débit.

Nous avons d'abord passé au NIRS l'ensemble des échantillons disponibles, pour ensuite sélectionner un sous-échantillon représentatif de la diversité spectrale et ainsi constituer un set de calibration. Des dosages biochimiques ont été réalisés au laboratoire sur ce set de calibration pour mesurer les caractères suivants : teneur en lignines, ratios entre les sous-unités H, G et S des lignines, teneur en glucose, xylose et en extractibles. Des modèles de calibration ont alors été établis pour chacun des caractères sur 4/5 des échantillons du set de calibration par régression PLS (Partial Least Squares) et ces modèles ont ensuite été validés sur le cinquième restant (FIGURE 6). Une fois validés, ils ont été utilisés sur l'ensemble des échantillons pour analyser leur variabilité phénotypique.

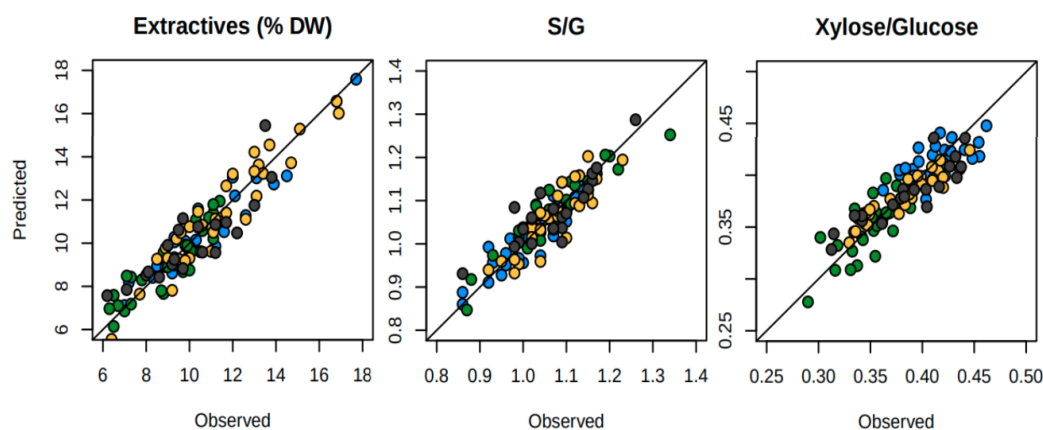


FIGURE 6 – Illustration de modèles de calibration NIRS développés pour le phénotypage à haut-débit des propriétés du bois

Valeurs observées et prédites pour le taux d'extractibles exprimé en pourcentage de la matière sèche (Extractives (% DM)) et les ratios entre sous-unités S et G des lignines (S/G) et entre Xylose et Glucose pariétaux (Xylose/Glucose). Les points colorés en bleu, vert et jaune correspondent aux échantillons du set d'entraînement collectés respectivement à Orléans en 2010, Savigliano en 2011 et Orléans en 2012. Les points colorés en gris foncé correspondent aux échantillons du set de validation.

Une large part de variabilité génétique a été mise en évidence pour tous les caractères étudiés. Des IGEs ont également été identifiées, mais leur variance était plus faible que la variance génotypique pour tous les caractères sauf les teneurs en extractibles et en glucose. Toutefois, une analyse fine de ces interactions a montré qu’elles étaient principalement dues à des changements de classement d’un faible nombre de génotypes, laissant ainsi une large gamme de matériel génétique stable et performant, d’intérêt pour l’amélioration génétique de la qualité du bois. Ces travaux ont été réalisés dans le cadre de la thèse de Mesfin Nigussie Gebreselassie que j’ai co-encadrée avec Catherine Bastien. Ils ont été valorisés par une publication scientifique dont je suis l’auteur de correspondance (Gebreselassie et al., 2017).

2.3.3.2 Pipeline de détection et de génotypage de SNPs à partir de séquences d’ARN

L’étape suivante, dans la mise en place du projet de recherche, a consisté à développer une pipeline pour la détection et le génotypage de SNPs à partir des séquences d’ARN. Ce pipeline a été développé sur les séquences produites dans le cadre d’une expérimentation pilote sur un sous-échantillon de 12 génotypes parmi ceux de la collection. Cette expérience pilote avait également pour objectif de bien préparer et valider chacune des étapes expérimentales depuis la collecte des échantillons jusqu’à l’analyse des séquences, afin qu’elle puisse ensuite être plus facilement étendue à un ensemble de génotypes plus large.

Si la technique RNAseq était déjà grandement utilisée pour les études de transcriptomique, relativement peu d’études avaient évalué son utilisation pour faire du génotypage lorsqu’elle est mise en œuvre sur une collection de génotypes. La stratégie d’analyse a consisté à déployer en parallèle plusieurs outils de détection et de génotypage de SNPs, pour ensuite déterminer quelle combinaison permettait d’obtenir le meilleur compromis entre nombre de SNPs et précision de génotypage. Pour cela, nous nous sommes notamment appuyés sur les données de génotypage par puce disponibles sur les génotypes de la collection (Faivre-Rampant et al., 2016) et avons montré que la combinaison de 3 outils était la modalité la plus pertinente. Cette modalité aboutissait à plus de 350 000 SNPs chez les 12 génotypes, ce qui permettait de densifier fortement les données de génotypage existantes pour les études de génétique d’association. Des premières analyses de structure de la diversité avec ces SNPs ont confirmé leur pertinence.

Ce travail a fait l’objet d’une publication scientifique dont je suis l’auteur de correspondance (Rogier et al., 2018).

2.3.3.3 Variabilité du transcriptome, réseaux de co-expression et prédiction

L’expérimentation pilote a permis de valider l’approche RNAseq qui a donc ensuite été déployée sur un ensemble de 240 génotypes, à raison de 2 répétitions par génotype, localisées dans 2 blocs du dispositif Orléanais. Concrètement, nous

avons échantillonné le xylème et le cambium de 480 arbres, nous avons extrait les ARN de ces 960 échantillons, puis nous avons reconstitué 480 pools équimolaires par arbre. Ces pools ont ensuite été envoyés en plateforme pour construction de banques d'ADNc et séquençage de type Illumina.

Après différentes étapes de pré-traitement des données de séquençage pour notamment prendre en compte et corriger les effets expérimentaux, nous avons construits des réseaux de co-expression à partir des corrélations génotypiques entre transcrits et avons identifiés 16 modules (groupes) de gènes. Nous avons ensuite étudié la connectivité au sein des modules pour distinguer les gènes centraux (ou cœurs) des gènes périphériques et étudier leurs caractéristiques. Nous avons notamment pu montrer que les gènes cœurs étaient fortement exprimés, très différenciés entre les populations d'origine tout en présentant des niveaux généralement faibles de variation génétique. A l'inverse, les gènes périphériques étaient faiblement exprimés, peu différenciés entre populations mais avec un niveau de variabilité génétique élevé. Ces caractéristiques suggèrent que les gènes cœurs sont plus contraints que les gènes périphériques, potentiellement par des phénomènes de sélection naturelle divergente entre populations. Malgré ces différences nettes entre caractéristiques des gènes cœurs et périphériques, d'un point de vue purement prédictif les différences entre sets de gènes étaient beaucoup moins claires, que l'on utilise un modèle purement additif comme la ridge regression ou un modèle supposé plus interactif comme les réseaux de neurones. Une approche de sélection de variables a tout de même montré un enrichissement significatif en gènes cœurs soulignant leur importance pour prédire des caractères, même si au final l'information qu'ils apportent n'est vraisemblablement pas suffisante pour atteindre de meilleurs niveaux de prédiction que ceux de gènes périphériques ou mêmes de gènes échantillonnés au hasard. Ces résultats renforcent l'idée d'une certaine redondance d'information telle que suggérée par la récente théorie omnigénique (Boyle et al., 2017).

Ce travail a fait l'objet du post-doctorat d'Aurélien Chateigner que j'ai encadré avec Leopoldo Sanchez et il a été valorisé par une publication scientifique dont je suis l'auteur de correspondance (Chateigner et al., 2020).

2.3.3.4 Génétique d'association à l'échelle du transcriptome

Le pipeline de détection et de génotypage à partir du RNAseq a également été déployé sur le jeu de données complet correspondant à 241 génotypes, ce qui a permis d'obtenir près de 875 000 SNPs détectés par au moins trois outils bioinformatiques et avec moins de 50% de données manquantes. La précision de génotypage pour les 3 841 positions communes avec la puce (Faivre-Rampant et al., 2016) était de l'ordre de 96%. Ces données issues de RNAseq et de puce ont alors été combinées et les données manquantes imputées, ce qui m'a permis d'aboutir à un jeu de données de près de 880 000 SNPs qui se retrouvent principalement dans la fraction génique puisqu'ils sont majoritairement issus du RNAseq. J'ai ensuite appliqué un filtre de 5% sur la fréquence de l'allèle minoritaire (MAF) puisque le pouvoir de détection des SNPs en génétique d'association est proportionnel au produit de la fréquence de

leurs allèles (qui détermine leur variance) et j'ai effectué des analyses de génétique d'association avec les 440 000 SNPs restant.

De façon notable pour la circonférence mesurée à Savigliano, des associations significatives ont été détectées sur la partie distale du chromosome 10 (FIGURE 7). Ces SNPs significatifs, au seuil de 5% après correction de Bonferroni, se retrouvent sur deux gènes adjacents, dont un, Potri.010G213000, est annoté comme une chalcone isomérase (CHI). Si on utilise le modèle MLM, seul le SNP le plus significatif ("top SNP"), situé dans le premier exon du gène CHI, est conservé et il ne reste plus de signal dans la région, ni ailleurs sur le génome. Ce top SNP est non-synonyme, c'est à dire qu'il conduit à un changement d'acide-aminé sur la protéine CHI, et il explique plus de 50% de la variabilité phénotypique observée entre génotypes pour la circonférence à Savigliano (sans prise en compte de la structure toutefois dans le calcul du pourcentage de variance expliqué). On peut noter que l'association existe aussi à Orléans mais l'effet du SNP sur le phénotype est moindre (20% de variance entre génotypes expliquée).

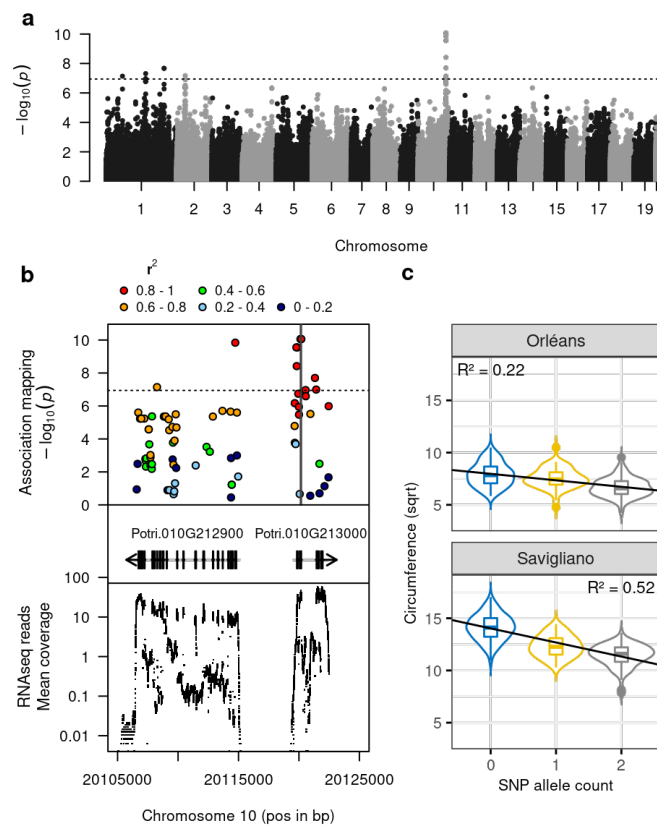


FIGURE 7 – Génétique d'association à l'échelle du transcriptome pour la circonférence mesurée à Savigliano

(a) Manhattan plot pour le scan d'association par modèle mixte exact simple locus. (b) Zoom sur la région significative du Manhattan plot avec coloration des SNPs selon leur niveau de déséquilibre de liaison avec le top SNP, représentation schématique des modèles de gènes, et représentation graphique de la couverture moyenne du RNAseq. (c) Illustration de l'effet du top SNP sur le caractère de circonférence mesuré à Orléans et Savigliano.

Pour aller plus loin dans l'interprétation de ce signal, nous avons tout d'abord utilisé les données d'expression des gènes. Ces données permettent d'identifier des corrélations négatives significatives entre l'expression des 2 gènes de la région et les phénotypes, et ces corrélations sont beaucoup plus fortes pour le gène CHI que pour l'autre gène ($R^2 = 0.35$ et 0.53 à Orléans et Savigliano pour CHI contre 0.18 et 0.07 à Orléans et Savigliano pour Potri.010G212900). Si on effectue par ailleurs une analyse d'association pour les expressions de chacun des 2 gènes (analyse eQTL), on retrouve du signal dans la région CHI (mécanisme de contrôle en CIS), et le patron de signal pour l'expression du gène CHI correspond à celui observé pour la circonférence tandis que ce n'est pas tout à fait le cas pour l'expression de Potri.010G212900. Ces résultats suggèrent que le gène CHI est un bon candidat pour le contrôle de la variabilité de la croissance dans notre collection.

J'ai ensuite étudié de façon plus précise le top SNP. De façon intéressante, ce SNP présente un très fort niveau de différenciation entre les populations d'origine, bien au-delà du 99^{ème} percentile de la distribution des F_{ST} de l'ensemble des SNPs (FIGURE 8). Cette forte différenciation entre populations se retrouve par ailleurs pour les caractères de croissance ainsi que pour l'expression du gène CHI.

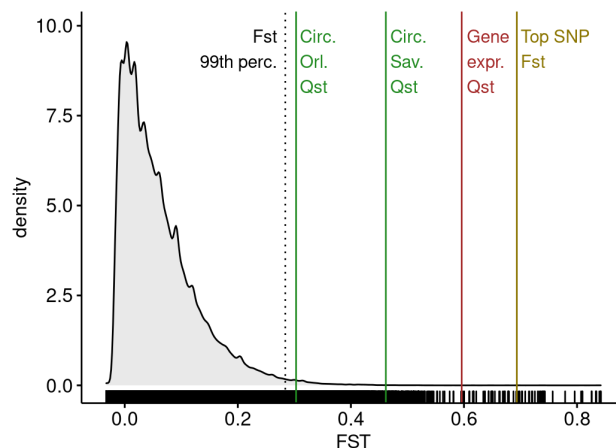


FIGURE 8 – **Indices de différenciation entre populations de peuplier noirs**
Densité de distribution des indices de différenciation F_{ST} entre populations d'origines de peupliers noirs estimés sur tous les SNPs disponibles. Les lignes verticales représentent les valeurs particulières suivantes : 99^{ème} percentile de la distribution de F_{ST} , indices de différenciation phénotypique Q_{ST} pour les caractères de circonférence et l'expression du gène CHI, et indices de différenciation du top SNP.

Ces résultats suggèrent que le top SNP contribue à la composante inter-populationnelle de la variabilité génétique du caractère de croissance, ce qui est d'ailleurs confirmé par une analyse par modèle linéaire mixte multivarié sur les deux sites et avec deux effets aléatoires du génotype, inter- et intra-populations (FIGURE 9). On peut voir par ailleurs dans cette analyse que le top SNP contribue aussi à la variabilité d'IGE à l'échelle inter-populationnelle, ce qui pouvait être attendu compte tenu du fait que son effet sur la croissance est moindre à Orléans par rapport à Savigliano.

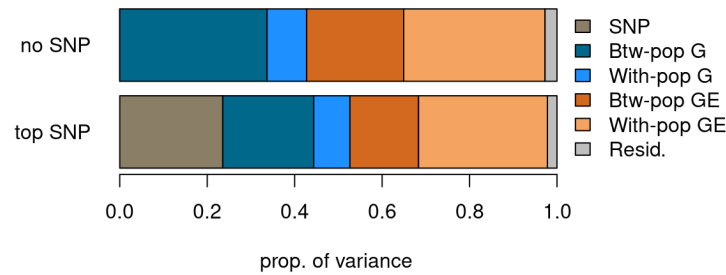


FIGURE 9 – **Effet du top SNP sur la partition de variance inter-sites**
 Partition de la variance entre génotypes pour la circonférence mesurée dans les deux sites en composantes inter- et intra-populationnelles génétique et d’IGE sans ou avec effet fixe du top SNP dans le modèle.

Pour illustrer ces résultats, j’ai alors représenté les corrélations entre phénotype, génotype et expression à l’échelle populationnelle (FIGURE 10). Dans le cas du phénotype et de l’expression du gène CHI, j’ai utilisé les moyennes par population, tandis que dans le cas du SNP, j’ai utilisé la fréquence de l’allèle alternatif. Comme c’était le cas à l’échelle génotypique, les résultats montrent que l’association est plus forte lorsque le caractère est mesuré à Savigliano et cela aussi bien pour le SNP que pour l’expression du gène CHI, bien que cette dernière ait été mesurée à Orléans. Cela peut s’expliquer par le fait que le caractère de croissance est plus exprimé à Savigliano qu’à Orléans à cause de conditions pédo-climatiques plus favorables.

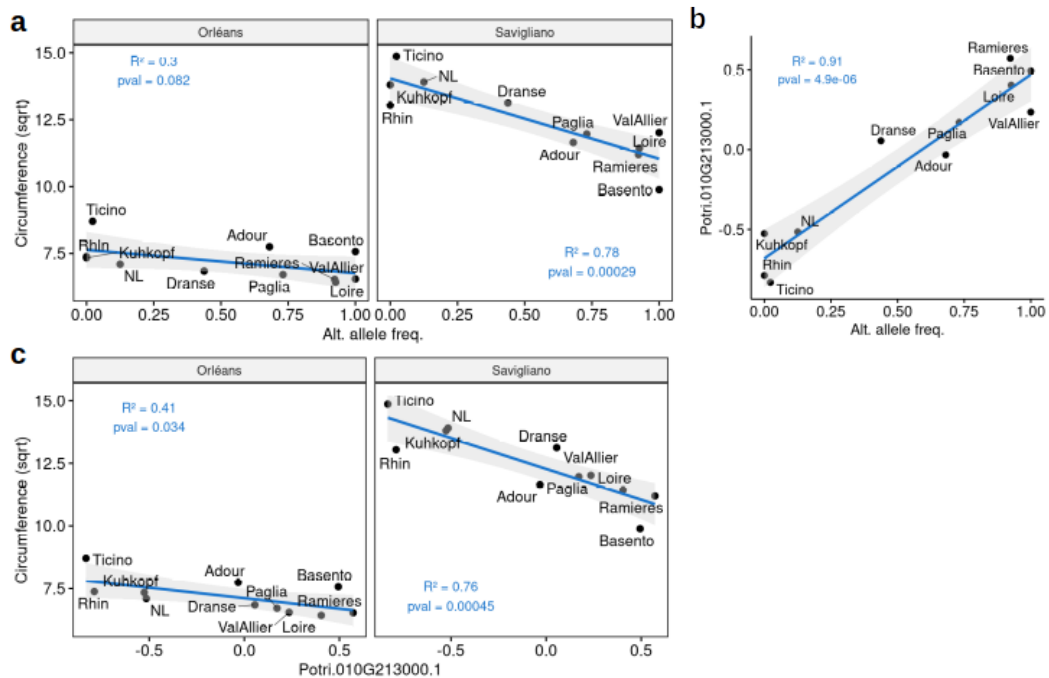


FIGURE 10 – **Illustration des associations à l’échelle populationnelle**
 Corrélations entre (a) phénotype et fréquence de l’allèle alternatif chez le top SNP, (b) phénotype et expression du gène CHI, et (c) expression du gène CHI et fréquence de l’allèle alternatif chez le top SNP.

Le patron observé pour le top SNP pourrait résulter d'un mécanisme de sélection naturelle divergente. En effet, il est quasiment fixé pour l'allèle de référence dans certaines populations plutôt localisés au Nord-Est de la zone géographique considérée, quasiment fixé pour l'allèle alternatif dans d'autres populations plutôt originaire du Sud et Sud-Ouest, et à fréquence intermédiaire dans quelques populations à l'interface entre les deux extrêmes comme Dranse et Paglia (FIGURE 11). Notre analyse montre que ce patron particulier explique la vaste majorité de la variabilité observée entre ces populations pour la croissance dans le site favorable et pour l'expression du gène CHI.

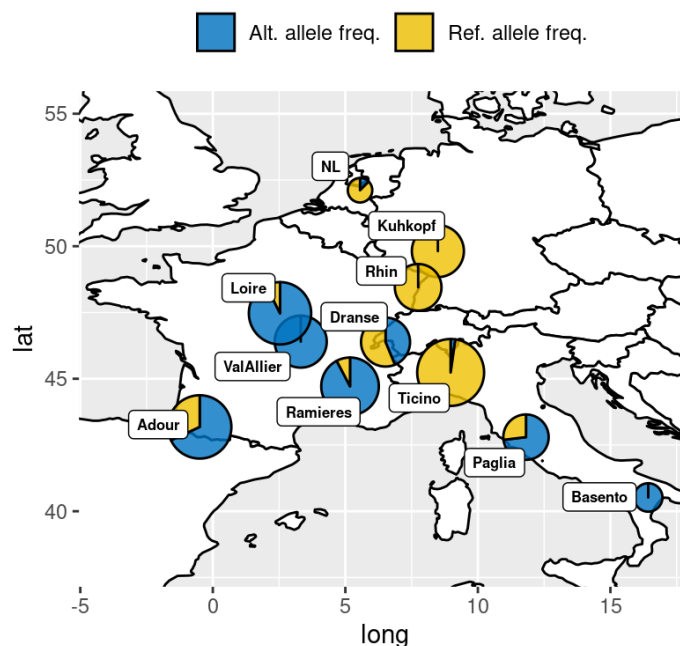


FIGURE 11 – Illustration de la différenciation génétique observée pour le top SNP.

Projection sur une carte des fréquences alléliques dans chacune des populations.

Étant donné que le gène CHI est connu et qu'il n'existe qu'en une seule copie dans le génome du peuplier (ce qui est plutôt rare), on peut légitimement se poser la question de l'interprétation de l'association détectée. En effet, ce gène est clé dans la voie de biosynthèse des flavonoïdes qui jouent des rôles importants dans la réponse des plantes aux stress biotiques et abiotiques (Jez and Noel, 2002). Aussi le phénomène observé pourrait témoigner d'un compromis entre croissance et défense. Par ailleurs, la voie de biosynthèse des flavonoïdes dérive des mêmes précurseurs que d'autres composés métaboliques secondaires comme les lignines qui composent les parois cellulaires, ce qui pourrait aussi expliquer l'association détectée. Toutefois, dans ce cas de figure, on aurait plutôt pu s'attendre à détecter cette association avec les caractères biochimiques comme la teneur ou la composition en lignines plutôt qu'avec la croissance, ce qui n'était pas le cas.

En ce qui concerne le mécanisme qui pourrait être mis en œuvre, on sait que le top SNP est non synonyme mais c'est le cas également de 3 autres SNPs localisés dans le gène CHI et qui étaient tous significatifs lors du scan d'association initial. Les données d'expression suggèrent que l'effet du SNP causal affecte l'expression du gène (régulation en CIS), qui impacte à son tour la croissance, avec une relation négative : les individus poussent d'autant plus que le gène est faiblement exprimé dans leurs tissus, ce qui renforce l'idée d'un compromis entre production de flavonoïdes et croissance. Pour aller plus loin dans l'identification du variant causal et dans la compréhension du mécanisme mis œuvre, une approche d'édition du génome ciblant les SNPs candidats pourrait être utile. Cela permettrait aussi de valider l'association car on est dans une situation assez inconfortable du point de vue de l'effet confondant de la structure. En effet, le locus détecté tout comme le caractère de croissance et l'expression du gène CHI sont fortement structurés. Toutefois, deux arguments supplémentaires supportant l'effet potentiel de ce locus sur la croissance peuvent être amenés :

- il correspond à un "QTL hotpot" précédemment identifié pour la production de biomasse dans une population F2 issue d'un croisement *P. trichocarpa* × *P. deltoïdes* (Rae et al., 2009) ;
- l'effet des allèles du locus est significatif et il va dans le même sens pour la croissance évaluée dans un autre site au sein d'un plan de croisement de peupliers noirs, utilisé par ailleurs pour développer des modèles de prédiction génomique (Pégard et al., 2020).

Dès l'obtention des premiers résultats, des transformations génétiques avaient été tout de même lancées afin de valider l'effet du gène CHI en collaboration avec Annabelle Déjardin et Gilles Pilate de l'UMR BioforA. Il s'agit de transformations de type RNAi sur l'hybride de peuplier tremble modèle (clone INRAE 717-1B4), afin notamment de produire des arbres sous-exprimant le gène et dont on s'attend d'après les analyses précédentes à ce qu'il croissent plus. L'évaluation des transformants est en cours afin de compléter et rendre les résultats plus robustes avant leur publication.

2.3.3.5 Intégration multi-omiques

Si un premier travail de prédiction de phénotypes avec les données transcriptomiques avait été réalisé dans le cadre du travail post-doctoral d'Aurélien Chateigner, le gain potentiel de les combiner aux SNPs dans un contexte prédictif restait à évaluer. Par ailleurs, des travaux de prédiction génomique dans des plans de croisements de peuplier noir venaient d'être réalisés au sein de l'unité, soulignant tout son intérêt pour le programme d'amélioration génétique (Pégard et al., 2020). Dans ce contexte, l'existence conjointe de données génomiques et transcriptomiques sur la même collection d'individus ouvrait des perspectives d'intégration des données multi-omiques pour améliorer la prédiction des phénotypes. C'est dans ce cadre que la thèse d'Abdou Rahmane Wade a été initiée avec le soutien du métaprogramme INRA Selgen (projet EPINET que j'ai coordonné) et du programme Européen de

la recherche H2020 (projet B4Est). Il s'agit d'une thèse en cours, que je co-encadre avec Leopoldo Sanchez et Harold Duruflé de l'UMR BioforA¹.

Le premier travail de cette thèse a consisté à combiner SNPs et transcrits par concaténation et à construire des modèles de prédiction des caractères basés sur la régression pénalisée. Abdou a ensuite comparé la capacité prédictive des modèles combinés par rapport à celle obtenue avec les SNPs ou les transcrits seuls. Une certaine variabilité a été observée selon le caractère considéré, et de façon notable le modèle combiné était plus fréquemment avantageux pour les caractères évalués à Orléans, c'est à dire dans le site où avait eu lieu l'échantillonnage pour la transcriptomique. Pour aller plus loin dans l'interprétation des différences de qualité de prédiction observées entre caractères et sites, nous avons entrepris une analyse eQTL qui permet de caractériser les relations entre les deux couches omiques de prédicteurs, SNPs et transcrits. Cette analyse a notamment mis en évidence un certain nombre de SNPs "hubs" associés en TRANS avec un grand nombre de transcrits (FIGURE 12).

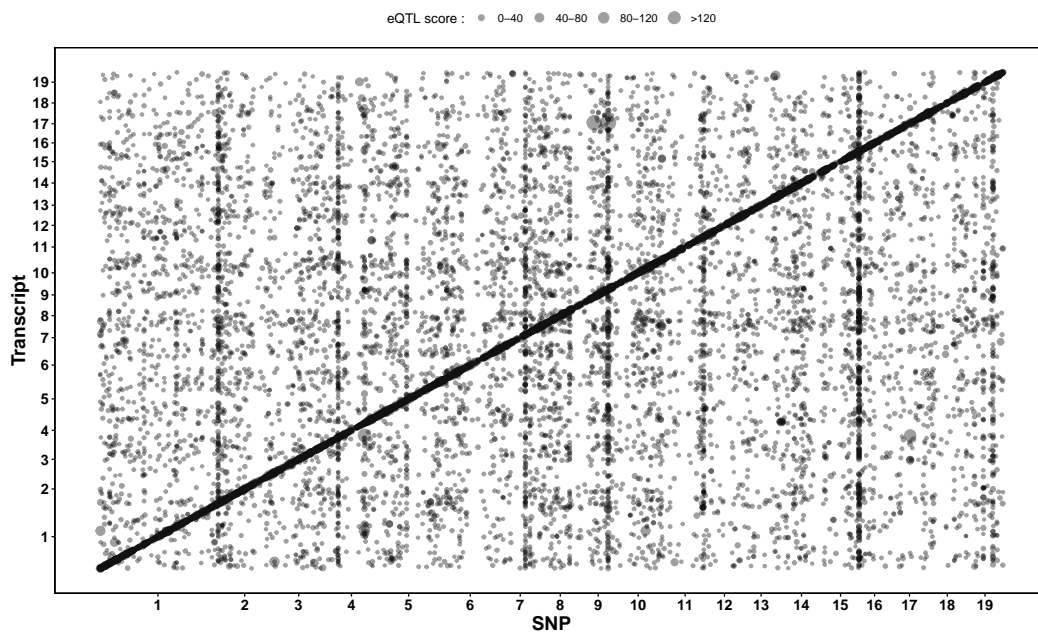


FIGURE 12 – Carte des eQTLs détectés

Les points représentent les associations significatives détectées entre chaque transcrit et chacun des SNPs, leur diamètre étant proportionnel au score d'association. Les points sont positionnés selon la localisation génomique des SNPs et transcrits auxquels ils correspondent.

A partir de cette analyse, les SNPs et transcrits ont été regroupés en trois catégories principales selon leurs relations identifiées lors de l'analyse eQTL : CIS, TRANS, pas d'association. Des corrélations négatives significatives ont alors été identifiées

1. Harold a rejoint l'encadrement de la thèse suite à ma mobilité et à son recrutement au sein de l'UMR BioForA.

entre le changement de rang des prédicteurs (qui témoigne de leur importance statistique) et l'avantage prédictif dans le modèle multi-omiques par rapport aux modèles simple-omique, pour les catégories correspondant aux eQTL TRANS, CIS et aux transcrits régulés en CIS pour les caractères évalués dans le site d'Orléans (FIGURE 13).

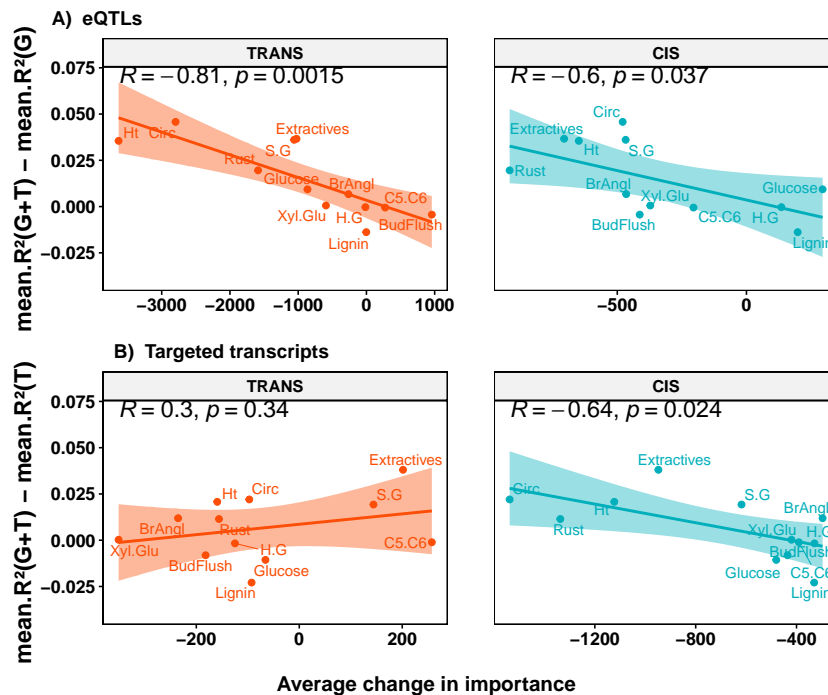


FIGURE 13 – Relation entre changement d'importance des prédicteurs et avantage prédictif en multi-omiques

Régressions sur les caractères mesurés à Orléans entre les changements d'importance des prédicteurs et le gain prédictif en modèle multi-omiques combinant SNPs et transcrits par rapport aux modèles simple-omique.

Ces résultats suggèrent que l'intégration bénéfique se produit lorsque la redondance des prédicteurs est diminuée, laissant la place à d'autres prédicteurs moins importants mais complémentaires. Ils sont par ailleurs renforcés par une analyse d'enrichissement en terme GO (Gene Ontology), puisque les prédicteurs associés à un gain de performance du modèle combiné multi-omiques sont enrichis pour des processus assez spécialisés en relation avec les caractères prédits.

Ces résultats soulignent l'effet de la redondance des prédicteurs dans le contexte multi-omiques dans le succès de l'intégration et ainsi ouvrent des perspectives pour le développement de nouveaux modèles permettant de mieux la prendre en compte et ainsi d'atteindre de meilleures performances prédictives. Ce travail fait l'objet d'une article soumis dont je suis l'auteur de correspondance et qui a été déposé sur bioRxiv (Wade et al., soumis).

2.3.4 Au-delà du transcriptome, implication de la méthylation de l'ADN dans la variabilité phénotypique

En parallèle de ce travail effectué dans le cadre du projet SYBIOPOP, j'ai développé une collaboration avec Stéphane Maury de l'Université d'Orléans sur l'épigénétique et plus spécifiquement sur la variabilité de la méthylation de l'ADN chez le peuplier en réponse au stress hydrique, mais également dans un contexte de diversité génétique (Lafon-Placette et al., 2018; Le Gac et al., 2018; Sow et al., 2018, 2021).

Cette collaboration s'est notamment poursuivie dans le cadre du projet EPI-TREE (ANR-17-CE32-0009) en cours, dans lequel nous sommes notamment en train d'ajouter une couche de données épigénétiques (méthylation de l'ADN) aux données génomiques, transcriptomiques et phénotypiques précédemment produites sur la collection de peuplier noir dans le cadre du projet SYBIOPOP. Dans un premier temps, nous avons généré le méthylome d'un sous échantillon de 20 génotypes représentatifs de la collection par séquençage après traitement au bisulfite de l'ADN des mêmes échantillons que ceux qui avaient été précédemment analysés par RNAseq. L'objectif principal de cette première expérience était d'identifier les régions les plus intéressantes à capturer dans un deuxième temps sur l'ensemble des 241 génotypes de la collection. Par ailleurs, les premières analyses de ces données ont pu mettre en évidence des différences notables selon le contexte de méthylation, les méthylations en contexte CG permettant de marquer relativement bien la structure de la collection, tandis que cette corrélation avec la structure génétique s'affaiblit en contexte CHG pour devenir quasiment inexistante en contexte CHH. La poursuite de ces analyses est en cours en collaboration avec les partenaires du projet qui sont notamment impliqués dans le work-package dédié à l'analyse et à l'intégration des données. En ce qui concerne la caractérisation de la méthylation de l'ADN des régions capturées sur l'intégralité de la collection, les données devraient être produites en 2022 et devrait apporter des informations sur les relations entre SNPs, méthylation de l'ADN, expression des gènes et caractères d'intérêt.

Si je suis encore co-responsable du workpackage dédié à la caractérisation de la méthylation de l'ADN dans la collection de peupliers, du fait de ma mobilité je ne pourrai pas autant m'impliquer dans leur analyse que si j'étais resté dans l'UMR BioForA. Odile Rogier, qui est ingénieur en bioinformatique au sein de l'unité, a pris le relais pour représenter l'unité au sein du projet et notamment gérer, au-delà des aspects scientifiques, les aspects financiers et administratifs. Par ailleurs, Harold Duruflé, qui a récemment été recruté comme chargé de recherche dans l'unité, participera aussi, en relation avec Odile, aux analyses des données générées dans le cadre de ce projet.

2.4 Prédiction phénotypique

Mon implication dans le phénotypage à haut-débit par NIRS, dans le cadre notamment du projet SYBIOPOP que j'ai coordonné, mais aussi du projet TOP-

WOOD (H2020-MSCA-RISE-2014) dont j'ai coordonné le work-package dédié à cette méthodologie, m'a permis d'acquérir de solides compétences dans ce domaine d'application. Certains travaux ont d'ailleurs fait l'objet de valorisations sous forme de publications scientifiques (Pulkka et al., 2016; Sergent et al., 2020). C'est dans ce contexte que j'ai développé le concept de prédiction phénotypique.

2.4.1 Définition et preuve de concept

La prédiction phénotypique propose d'utiliser la signature spectrale des échantillons à la place des marqueurs moléculaires pour inférer des ressemblances entre individus et prédire des caractères d'intérêt. Initialement, le terme de sélection phénotypique avait été proposé par analogie avec la sélection génomique (Rincent et al., 2018), même si comme cela est souvent le cas aussi en génomique, on se contente généralement d'effectuer des prédictions qui peuvent ensuite être utilisées à des fins de sélection dans les programmes d'amélioration. Ces travaux ont été effectués en collaboration avec Renaud Rincent (UMR GDEC puis GQE) qui travaillait sur blé, ce qui permettait notamment de généraliser l'approche au-delà du peuplier.

L'hypothèse principale derrière le concept de prédiction phénotypique est la suivante : les données NIRS témoignent principalement des propriétés physico-chimiques des échantillons analysés qui sont elles-mêmes au moins partiellement sous contrôle génétique. Alors, les spectres peuvent capturer une part de variabilité génétique et ainsi être utiles pour inférer des matrices de ressemblances entre génotypes directement utilisables en prédiction dans un équivalent phénotypique du G-BLUP. Comme le coût de préparation des échantillons et d'obtention des spectres est relativement limité et le débit de la méthode est par ailleurs assez élevé, ces avantages justifiaient l'approche de prédiction phénotypique, et cela même dans des situations plutôt avantageuses pour la prédiction génomique dans lesquelles les coûts de génotypage étaient déjà assez limités (par exemple chez certaines espèces phares de grandes cultures).

Conceptuellement, la prédiction phénotypique implique plusieurs différences par rapport à l'utilisation classique de la NIRS en phénotypage à haut-débit :

- Les modèles sont établis à l'échelle du génotype et non pas du tissu analysé. En amélioration des plantes comme on travaille généralement avec des répétitions des génotypes, cela implique de passer par une étape de calcul du spectre moyen par génotype.
- Être à l'échelle du génotype offre alors la possibilité d'effectuer des calibrations entre environnements, c'est à dire qu'avec des spectres collectés sur des plantes dans un environnement donné il est possible d'entraîner un modèle avec des phénotypes collectés sur les mêmes génotypes mais dans des conditions environnementales différentes (scénario 2 dans Rincent et al., 2018).
- D'ailleurs pour que cela fonctionne, les caractères à prédire n'ont pas nécessairement besoin d'être liés aux propriétés chimiques des tissus analysés par NIRS. C'est le cas par exemple lorsque l'on prédit le rendement en grain avec des spectres collectés sur feuilles.

Pour réaliser la preuve de concept, nous avons donc tout d'abord partitionné le long du spectre la variabilité observée en variances génétique, d'IGE et résiduelle au moyen de modèles linéaires mixtes multivariés. Ces analyses ont montré que la variabilité génétique pouvait représenter jusqu'à 60% de la variabilité phénotypique pour certaines longueurs d'onde dans le cas des spectres collectés sur grains de blés. Tandis que sur feuilles de blé et bois de peuplier la part de variance attribuée à la génétique pouvait atteindre jusqu'à 40% de la variance totale. De plus selon le tissu considéré, une part de variance d'IGE non négligeable pouvait être également captée par le spectre. Ces résultats confirmaient bien l'hypothèse initiale que les spectres NIRS incluent une certaine information génétique. L'étape d'après consistait alors à tester si cette information était suffisante pour prédire diverses catégories de caractères non nécessairement liés aux propriétés physico-chimiques des tissus sur lesquels les spectres avaient été collectés.

Nous avons donc mis en œuvre et comparé prédictions phénotypiques et génomiques sur blé et peuplier pour différents caractères d'intérêt. Les résultats de prédictions à partir de spectres collectés sur grains de blés étaient très spectaculaires avec des capacités prédictives des modèles systématiquement supérieures ou égales à celles obtenues avec plus de 80 000 SNPs pour la date de floraison et le rendement, et cela même dans cas du scénario S2 pour lequel les spectres et caractères prédits ne provenaient pas des mêmes environnements (FIGURE 14). Avec les spectres collectés sur feuilles, les précisions de prédictions phénotypiques étaient meilleures que les précisions de prédiction génomique pour la date de floraison, et de qualité comparable pour le rendement. Chez le peuplier les résultats étaient plus variables, avec de bonnes qualités de prédiction pour les caractères de croissance, des résultats plus mitigés pour la date de débourrement, et des qualités intermédiaires pour la date d'arrêt de croissance et la résistance à la rouille foliaire.

En tirant partie de données issues de plusieurs dispositifs expérimentaux sur blé, nous avons par ailleurs pu montrer que les précisions de prédiction phénotypique pour le rendement pouvait être de bonne qualité même dans le cas de figure où le rendement prédit dans un site donné n'était pas corrélé au rendement des plantes localisées sur un autre site et sur lesquelles les spectres avaient été collectés. Ce résultat renforce l'idée que les prédictions phénotypiques utilisent bien une information génétique plutôt que des corrélations indirectes entre phénotypes.

Enfin, nous avons utilisé des simulations basées sur des scénarios de qualité de prédiction et de coûts pour l'acquisition de données NIRS et le génotypage, afin d'identifier dans quelles situations la prédiction phénotypique pouvait présenter un avantage en terme de gain génétique attendu par rapport à de la prédiction génomique. Dans tous les cas de figure, la prédiction phénotypique aboutissait à un gain par rapport à la prédiction génomique et ce gain pouvait aller, dans les situations les plus favorables, jusqu'à 130%. Nous avons ensuite calculé les gains génétiques attendus, sur blé et peuplier, sachant les précisions de prédictions précédemment obtenues et une estimation des coûts d'acquisition des données. Chez le blé, l'augmentation de gain génétique attendu était presque toujours positive et il pouvait aller jusqu'à 222% dans le cas le plus favorable pour le rendement. Chez le peuplier,

comme dans le cas des précisions de prédiction, les résultats variaient en fonction du caractère considéré, les gains étant majoritairement positifs pour la croissance, et intermédiaires pour la date d'arrêt de croissance et la résistance à la rouille.

Ce travail permettait donc de faire la preuve du concept et ouvrait de nombreuses perspectives d'application dans le contexte de l'amélioration génétique. On peut d'ailleurs souligner quelques d'applications récentes, avec notamment des données de réflectance obtenues par imagerie aéroportée, chez diverses espèces de céréales (Krause et al., 2019; Galán et al., 2020; Lane et al., 2020; Zhu et al., 2021).

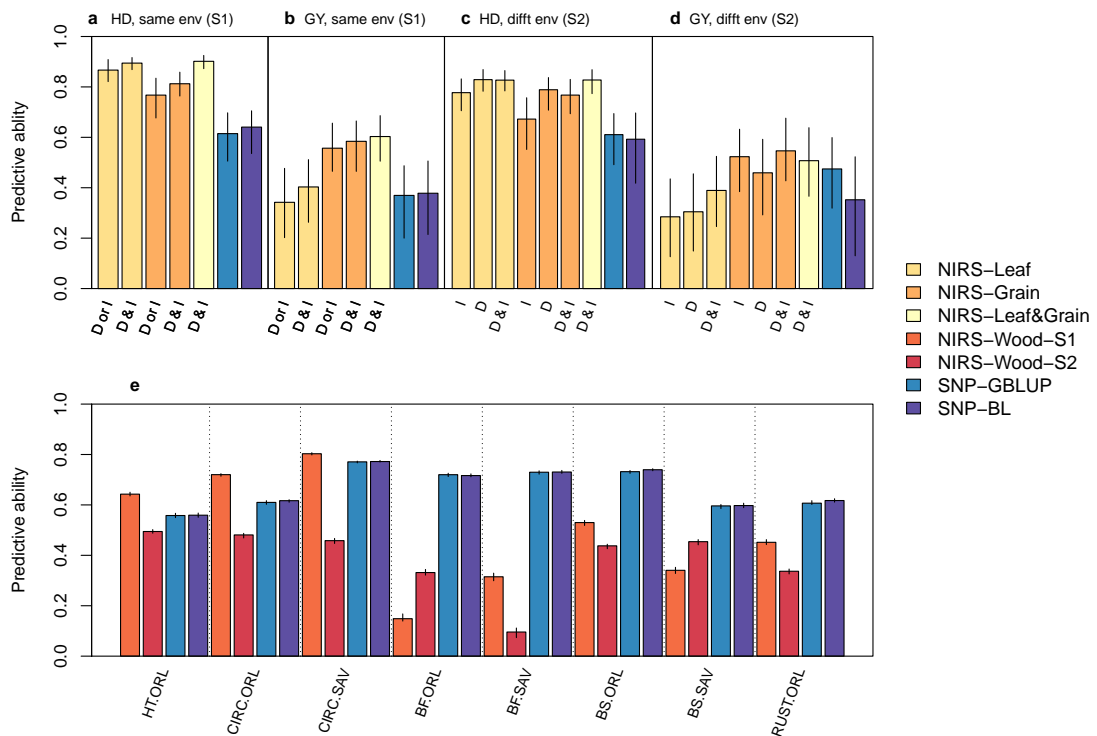


FIGURE 14 – Précisions des prédictions phénomique et génomique chez le blé et le peuplier.

Précisions des prédictions obtenues par validation croisée avec SNPs ou spectres NIRS collectés sur feuilles ou grains de blé (a-d) et bois de peuplier (e). Deux scénarios ont été considérés : S1 dans lequel les spectres et caractères proviennent du même environnement (a, b, e) et S2 dans lequel les spectres et caractères proviennent d'environnements différents (c, d, e). Les caractères prédits chez le blé sont la date de floraison (a, c) et le rendement (b, d). Les barres dans les graphiques a, b, c et d sont annotées selon les conditions de culture des plantes sur lesquelles les spectres ont été collectés (I : irrigué, D : stress hydrique). Les barres dans le graphique e sont annotés avec la combinaison caractère - site (HT : hauteur, CIRC : circonférence, BF : débourrement, BS : arrêt de croissance, RUST : résistance à la rouille foliaire, ORL : site d'Orléans, SAV : site de Savignano, Italie).

2.4.2 Application chez la vigne

Les premiers travaux que j'ai effectués après ma mobilité au sein de l'équipe DAAV de l'UMR AGAP Institut concernent cette thématique de sélection phénomique. L'objectif était de tester cette nouvelle méthodologie sur vigne à partir de spectres collectés sur bois et feuilles, et de la comparer à la sélection génomique en cours d'évaluation par ailleurs. Ces travaux ont été effectués dans le cadre de la thèse de Charlotte Brault que j'ai co-encadrée avec Agnès Doligez, Loïc Le Cunff, Timothée Flutre, et Patrice This¹.

Outre l'intérêt de tester la prédiction phénomique sur une nouvelle espèce, notamment pérenne, ce travail avait pour objet d'apporter des informations sur la façon dont fonctionne la prédiction phénomique et notamment sur les facteurs qui pourraient influencer sa précision.

Pour répondre à ces enjeux, nous avons défini un protocole permettant de collecter du matériel pour les prises de spectres sur un très grand nombre d'individus et avons appliqué ce protocole dans deux dispositifs complémentaires de l'équipe, implantés au domaine du Chapitre à Villeneuve-lès-Maguelone (34). Ces dispositifs correspondent à un plan de croisement demi-diallèle avec 10 croisements issus de 5 parents (Tello et al., 2019) et un panel de diversité qui capture la variabilité génétique de la vigne cultivée (Nicolas et al., 2016). Des échantillonnages de bois et de feuilles ont été réalisés au cours de deux années successives sur ces 2 dispositifs, afin notamment d'évaluer les effets de l'année et du tissu sur la qualité des prédictions phénomiques. Les spectres NIRS ont été acquis au laboratoire en collaboration avec Martin Ecartot (équipe GE²pop, UMR AGAP Institut). En ce qui concerne les caractères cibles, nous avons profité de données précédemment collectées sur ces dispositifs par l'équipe. Ces caractères correspondent aux catégories suivantes : phénologie, vigueur, production, morphologie de la grappe, et composition en acides des baies à maturité.

Dans un premier temps, nous avons décomposé la variabilité phénotypique le long du spectre en variabilité entre croisements ou structure génétique (selon le dispositif), génotypes, années, tissus, leurs interactions ainsi que celle due à des effets expérimentaux. Un certain niveau de variabilité entre croisements ou groupes génétiques, ainsi qu'entre génotypes a pu être mis en évidence lorsque l'on considère les tissus séparés. En revanche, mélanger les données collectées sur bois et feuilles conduit à une forte réduction de la variabilité génétique au profit de leurs interactions avec l'effet tissu. Cela n'est pas tout à fait surprenant, étant donné que les spectres collectés sur bois et feuilles sont assez différents puisque ces tissus n'ont pas la même structure physique ni la même composition chimique. Pour aller plus loin dans l'analyse de la part de signal génétique capturé par les spectres, nous avons effectué des analyses de co-inertie entre matrices de spectres collectés sur feuilles, bois et matrice de données génotypiques. Ces analyses mesurent la ressemblance entre paires de matrices de données NIRS et/ou génotypiques. Elles confirment que les spectres

1. J'ai rejoint l'encadrement de la thèse à environ mi-parcours, suite à ma mobilité et à celle de Timothée Flutre vers l'UMR GQE.

peuvent capturer un certain niveau de variabilité génétique avec des coefficients de concordance entre matrices de spectres et de SNPs atteignant 0.59 sur le demi-diallele. Les effets croisement ou structure génétique (selon le dispositif) semblent grandement contribuer à cette observation, puisque lorsque ces effets n'étaient pas pris en compte dans le calcul, les co-inerties entre spectres et SNPs étaient toujours plus faibles. De façon surprenante et notable, les co-inerties calculées entre les 2 matrices de spectres (bois *vs.* feuilles) étaient plus faibles qu'entre spectres (bois ou feuilles) et SNPs, et cela, quels que soient le tissu ou le dispositif considérés. Ces résultats confirment les observations précédemment réalisées sur blé et peuplier, qui montraient un certain niveau de variabilité génétique le long du spectre, mais cette fois-ci en utilisant une approche plus globale basée sur les matrices dans leur ensemble.

Nous avons ensuite fait des prédictions des caractères, et avons d'abord montré que d'extraire des valeurs génotypiques prédites (BLUPs) le long du spectre lors de l'étape de décomposition de la variance permettait d'augmenter systématiquement les précisions de prédiction par rapport à des moyennes génotypiques. Deux facteurs peuvent expliquer cela : (i) l'utilisation de modèles linéaires mixtes le long du spectre permet de prendre en compte et corriger certains effets potentiellement indésirables comme les effets expérimentaux, et (ii) la prise en compte explicite d'un effet croisement ou structure génétique dans les modèles a permis de se rapprocher de la matrice de données génotypiques comme souligné par l'analyse de co-inertie. Nous avons donc ensuite utilisé les prédictions basées sur les BLUPs pour comparer prédictions phénotypiques entre tissus et années. De faibles différences de qualité de prédiction phénotypique ont été observées selon ces deux modalités, même si l'intégration des deux années via les BLUPs et la combinaison des deux tissus de façon additive sous la forme de deux effets aléatoires dans le modèle semblait être globalement la situation la plus favorable. Ainsi, cette dernière modalité a été conservée pour la comparaison avec les prédictions génomiques. Si globalement une meilleure précision de prédiction a été obtenue avec la génomique par rapport à la NIRS, pour certains caractères comme la vigueur, le poids de baies ou la date de maturité, la prédiction phénotypique semble tout à fait satisfaisante et cela dans les deux dispositifs. On peut aussi noter que les précisions des prédictions phénotypique et génomique étaient corrélées, avec une pente et intercepts d'ordre de grandeur similaire entre les deux dispositifs (FIGURE 15). Cette corrélation suggère que des facteurs communs influencent les précisions de prédiction génomique et phénotypique, ce qui en quelque sorte confirme que les prédictions phénotypiques passent bien par de la génétique et semblent ainsi assez généralisables à d'autres caractères notamment.

On peut également ici revenir sur le fait que les caractères prédits avaient été évalués un certain nombre d'années (3 à 10 ans) avant les prises de spectres. L'évaluation des caractères sur au moins deux années et les prises de spectres sur deux années également (bien que différentes), ainsi que la modélisation et l'utilisation des BLUPs, permettent de minimiser les effets millésimes qui sont connus pour influencer les rendements et la qualité des récoltes en viticulture. On sait également que les spectres sont assez sensibles aux conditions environnementales et de fait on

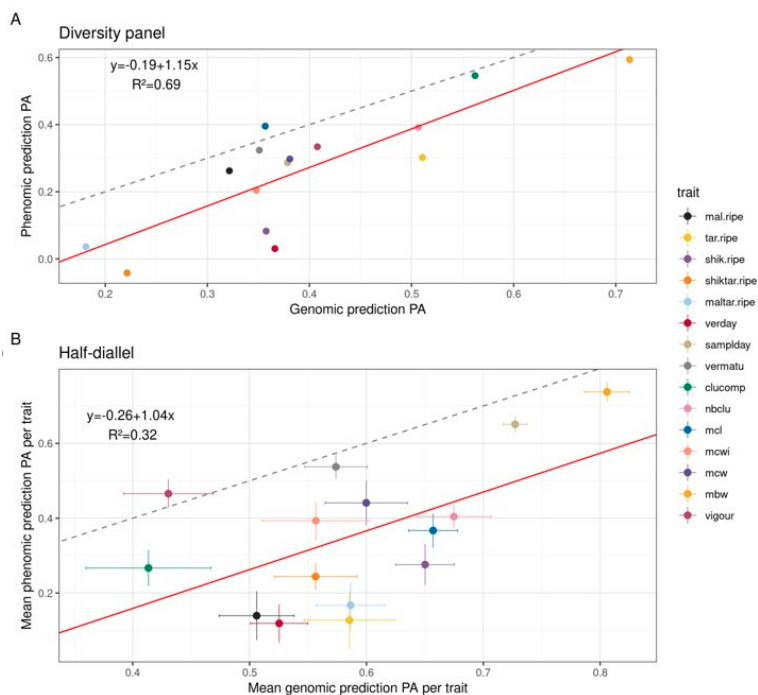


FIGURE 15 – Précisions des prédictions phénomique et génomique chez la vigne.

Comparaison des précisions de prédiction phénomique et génomique sur 15 caractères au sein de deux dispositifs : un panel de diversité (A) et un plan de croisement demi-diallèle (B). Dans le demi-diallèle, les valeurs correspondent aux précisions de prédiction moyennes sur les 10 croisements. La droite rouge représente la droite de régression correspondant à l'équation, tandis que la droite noire en pointillés représente la droite d'identité.

peut supposer qu'ils pourraient contribuer à prédire ces effets. Pour augmenter la précision des prédictions, il aurait fallu que les spectres soient collectés au cours des mêmes années que celles pour lesquelles les caractères ont été évalués. Toutefois, dans une perspective d'utilisation en sélection, l'intérêt résiderait plutôt dans une prise de spectres précoce en phase juvénile. En revanche, la capacité des spectres à capturer d'éventuels effets environnementaux et notamment des IGEs pourrait tout à fait être pertinente dans un contexte d'évaluation multi-sites, je reviendrai sur ces deux points dans la partie perspectives.

Ce travail fait l'objet d'un article en cours de finalisation dont je serai l'auteur de correspondance.

2.5 Conclusions sur la partie bilan

Ce bilan, qui retrace mes activités scientifiques depuis environ 15 ans, met en évidence un certain nombre de contributions à la recherche en génétique et amélioration des plantes. Il s'agit notamment de méthodologies innovantes pour la génétique

d'association et pour la prédiction de caractères d'intérêt à la fois via les prédictions génomique et phénomique et l'intégration multi-omiques. J'ai par ailleurs, dans le cadre du projet SYBIOPOP, contribué à la compréhension de mécanismes impliqués dans la production de biomasse chez le peuplier avec une approche originale de biologie intégrative. Dans ce projet, j'ai notamment produit de gros jeux de données¹ qui font encore l'objet de travaux et de projets de recherche en cours. Ces travaux étaient relativement lourds et ils ont nécessité la coordination d'un certain nombre de personnels permanents et non permanents. J'ai donc pu acquérir une certaine autonomie dans la coordination d'expérimentations d'envergures. En outre, j'ai acquis de l'expérience dans la mise en œuvre de la recherche sur projet depuis la réponse aux appels d'offre, jusqu'à la valorisation des résultats, en passant par la conduite de projet et les aspects de gestion administrative, financière et de ressources humaines.

Pour ce qui est spécifiquement de la prédiction phénomique, avec Renaud Rincant, nous avons initié un réseau à l'échelle nationale afin de promouvoir son utilisation sur de nombreuses espèces d'intérêt végétales ou animales. Un des objectifs de ce réseau serait de proposer de nouveaux projets sur cette thématique récente, afin notamment de progresser sur la compréhension de facteurs qui affectent son application. Cette thématique a aussi facilité ma récente mobilité, en faisant une bonne transition entre mes travaux précédents sur peuplier et ceux que je réalise désormais sur la vigne. Cela s'est notamment manifesté par mon implication dans l'encadrement de la thèse de Charlotte Brault, dont un des trois chapitre porte spécifiquement sur la prédiction phénomique.

Enfin, bien que cela ne transparaît pas forcément dans le bilan scientifique, il me semble important de mentionner ici mon implication dans des activités de formation au-delà de l'encadrement, via notamment des écoles chercheurs sur les méthodes développées en génétique d'association, des formations en interne sur l'utilisation basique ou même avancée du logiciel R et de l'environnement RStudio, de l'enseignement en biostatistiques à l'Université d'Orléans, et de l'enseignement en génétique quantitative à Montpellier SupAgro et à l'Université de Montpellier. J'ai également participé à de nombreux comités de thèse, partageant ainsi mes compétences en génétique quantitative. Certains ont même débouché sur des collaborations productives.

1. Les données de séquençage ont été déposées dans des bases de données internationales.

Chapitre 3

Perspectives¹

Mes perspectives de travail concernent des projets mis en œuvre sur la vigne dans le cadre de ma nouvelle affectation au sein de l'équipe DAAV de l'UMR AGAP Institut. Ces travaux s'organisent en deux parties complémentaires qui correspondent aux axes de recherche de l'équipe intitulés : "identification des bases génétiques et moléculaires de caractères d'intérêt et de l'adaptation" et "intégration pour la prédiction des caractères et l'innovation variétale" (FIGURE 16). Dans ces deux parties complémentaires, j'accorde une certaine importance à l'adaptation aux contraintes abiotiques imposées par le changement climatique.

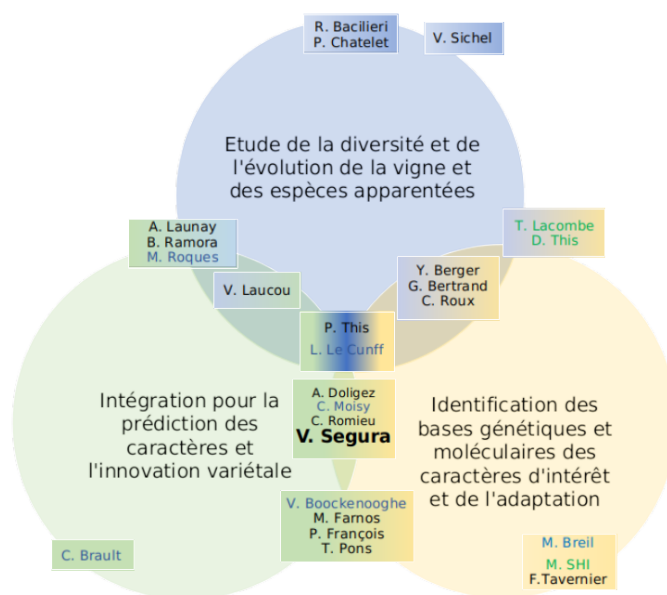


FIGURE 16 – Organigramme de l'équipe DAAV.

Les cercles représentent les trois axes de l'équipe, et les cadres représentent les agents et sont positionnés selon leur contribution aux différents axes. Les agents dont le nom est écrit en noir sont des personnels INRAE, ceux dont le nom est en bleu sont des personnels IFV (Institut Français de la Vigne et du Vin), tandis que ceux qui sont en vert sont des personnels de Montpellier SupAgro.

1. Les citations qui correspondent à des publications dont je suis co-auteur sont indiquées en gras dans le texte.

3.1 Contexte, enjeux et objectifs

Avec une production annuelle de 75.5 millions de tonnes de raisins frais sur 7.4 millions d'hectares, la viticulture constitue l'une des productions fruitières les plus importantes au monde (<https://www.oiv.int/fr/statistiques>). C'est également une culture qui occupe une place particulièrement importante en France, avec 6.25 millions de tonnes de raisins frais produits sur 786 000 hectares. La viticulture doit faire face de nos jours à deux défis majeurs : la réduction de l'utilisation de produits phytosanitaires et l'adaptation au changement climatique. L'innovation variétale par croisements est un des leviers permettant de répondre à ces défis.

En ce qui concerne la réduction de l'utilisation de produits phytosanitaires, les programmes d'amélioration génétique ont pour objectif l'introgession de gènes de résistance aux principales maladies de la vigne que sont le mildiou et l'oïdium. Ces maladies, provoquées par des champignons pathogènes originaires d'Amérique du Nord, ont été importées en Europe à la fin du 19^{ème} siècle. Ainsi, les résistances naturelles à ces pathogènes proviennent d'espèces sauvages du genre *Vitis* nord-américaines et donc ayant co-évolué avec les pathogènes. Ces résistances sont généralement de nature oligogénique avec des gènes sous-jacents à effet relativement fort. Aussi, pour anticiper et limiter d'éventuels contournements, les programmes génétiques visent l'introgession de plusieurs gènes de résistance (pyramidage). Actuellement en France, ce pyramidage concerne 3 loci pour chacune des deux maladies (mildiou et oïdium), qui sont sélectionnés par SAM (Sélection Assistée par Marqueurs) au cours du premier stade du programme d'amélioration (FIGURE 17).

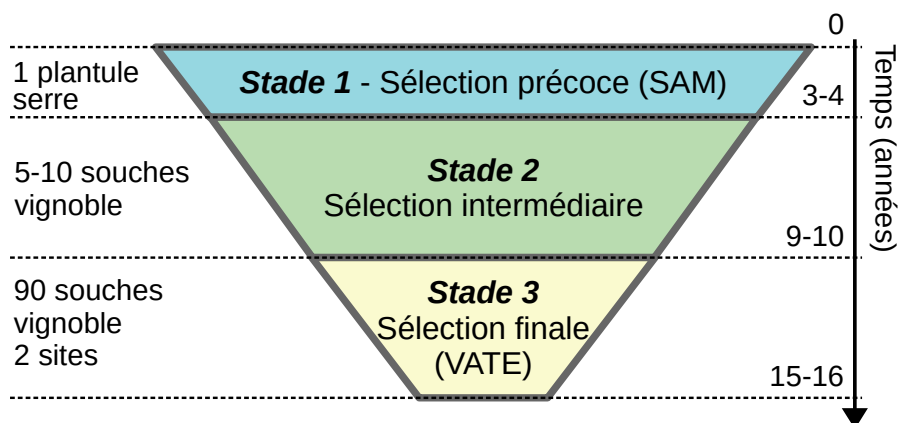


FIGURE 17 – Schéma de sélection de la vigne en France.

Le schéma de sélection de la vigne en France comprend actuellement 3 stades successifs. La pyramide inversée symbolise l'intensité de sélection à chaque étape qui est de l'ordre de 7%, ce qui signifie qu'il faut partir d'environ 3000 pépins pour obtenir une variété. (Tiré et adapté du projet scientifique de l'Unité Mixte Technologique Géno-Vigne®3).

Au cours du stade suivant, l'enjeu réside dans la sélection de génotypes avec de bonnes propriétés viticoles et œnologiques. Cette étape, qui dure environ 6 ans, est particulièrement lourde puisqu'elle consiste en des évaluations au vignoble sur

5 à 10 souches par génotype. La dernière étape du programme est une évaluation VATE (Valeur Agronomique, Technologique et Environnementale) sur 90 souches au vignoble et sur au moins deux sites, qui s'accompagne d'une description DHS (Distinction, Homogénéité, Stabilité) dans la collection de référence (Conservatoire des Ressources Biologiques Vigne de Vassal-Montpellier). Cette dernière étape est indispensable à l'inscription au catalogue. Compte tenu des contraintes et particularités de chacune de ces trois étapes, c'est clairement au stade de sélection intermédiaire que de nouveaux outils pour l'amélioration génétique, comme le phénotypage à haut-débit ou les prédictions génomique et phénomique, pourraient être déployées pour optimiser et accélérer les programmes de sélection. On pourrait imaginer *in fine* que la sélection intermédiaire se fasse sur la base de l'information génomique et/ou phénomique dès le premier stade, ce qui permettrait de gagner 6 ans sur les 15 à 16 années du programme actuel et ainsi de répondre plus vite aux attentes de la profession et des consommateurs. Je développerai spécifiquement mes perspectives concernant ces approches dans le chapitre 3.3.

En ce qui concerne l'adaptation aux contraintes induites par le changement climatique, les efforts notamment déployés lors des deux phases du projet pluridisciplinaire LACCAVE ont permis de caractériser de nombreux effets du changement climatique sur la viticulture et l'œnologie (<https://www6.inrae.fr/laccave>). Du point de vue climatique, les changements observés et prédits incluent notamment une hausse globale de la température et de la concentration atmosphérique en CO₂, des dérèglements aux niveaux des précipitations avec, globalement pour la France, une hausse au nord et une diminution au sud (GIEC, 2013). Ces changements influencent plusieurs processus physiologiques au cours du cycle de développement de la vigne, avec des conséquences sur le rendement et la qualité de la production (FIGURE 18, Ollat and Touzard, 2020).

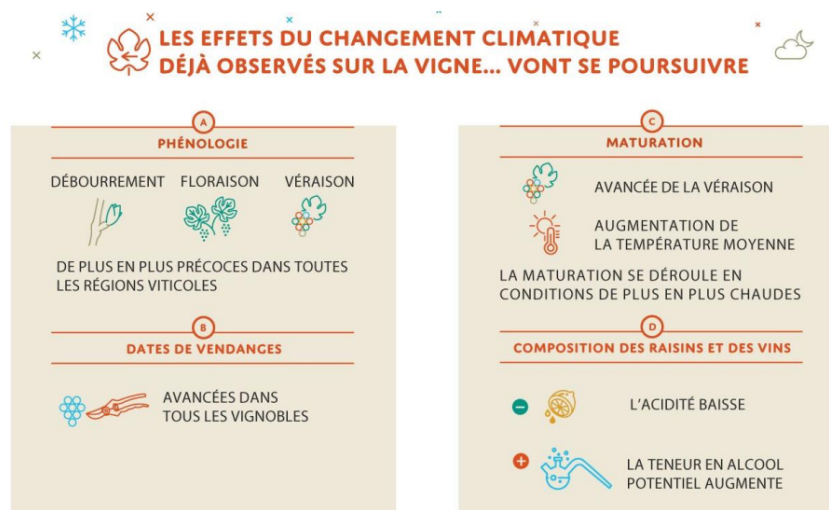


FIGURE 18 – Impacts du changement climatique sur la vigne et le vin. Infographie du projet LACCAVE (<https://www6.inrae.fr/laccave>).

En premier lieu, l'augmentation de la température affecte la phénologie de la vigne, ce qui se manifeste par une avancée dans le temps de ses principaux stades de développement : débourrement, floraison, véraison¹ et maturité. Cette avancée dans le temps combinée à la hausse globale des température fait que la période de maturation du raisin (après la véraison) se déroule dans des conditions de température plus élevées, qui affectent la composition des raisins et des vins avec notamment une augmentation de l'alcool potentiel et une diminution de l'acidité (Duchêne, 2016).

Si l'augmentation du CO₂ atmosphérique et de la température ont des effets plutôt positifs sur la production de biomasse via une augmentation de l'activité photosynthétique, ces effets sont à moduler en fonction des possibles contraintes hydriques (Duchêne, 2016). En effet, ces contraintes peuvent impacter le rendement différemment selon leur époque d'occurrence, sachant que par ailleurs le rendement se construit sur deux cycles de développement successifs, depuis l'initiation des primordia floraux jusqu'au remplissage des baies. La question de l'effet d'un stress hydrique ayant eu lieu l'année précédent la récolte est d'ailleurs peu documentée et originale. Elle est notamment au cœur du projet G2WAS (ANR-19-CE20-0024), que je présenterai plus en détails dans le chapitre 3.2.2. Pour ce qui est des effets plus directs d'une contrainte hydrique sur l'élaboration du rendement, une réponse instantanée au stress via la fermeture des stomates conduit à une économie d'eau au prix d'une réduction de la photosynthèse et de bénéfices induits par la transpiration sur le rafraîchissement des feuilles (Simonneau et al., 2014; Chaves et al., 2016). Des génotypes pouvant économiser de l'eau tout en conservant une certaine activité photosynthétique pourraient présenter un avantage en cas de contrainte hydrique puisqu'ils seraient en mesure de maintenir un certain niveau de rendement. La réduction de la transpiration nocturne pourrait partiellement contribuer à cet objectif d'augmentation de l'efficacité d'utilisation de l'eau (Coupel-Ledru et al., 2016). Toutefois, la mesure de ces caractères de fonctionnement au vignoble sur de nombreux individus est un verrou important qu'il est crucial de lever pour pouvoir les utiliser dans les schémas d'amélioration génétique (Gambetta et al., 2020).

En ce qui concerne la qualité de la vendange, l'instauration progressive d'un stress pendant la phase de maturation du raisin est généralement considérée comme un facteur favorable, à condition que celui-ci demeure modéré (Gambetta et al., 2020). Un stress marqué peut en revanche avoir des conséquences néfastes comme le flétrissement des baies vertes, ou affecter la composition en diverses molécules du métabolisme central (saccharose, acides aminés et organiques), ainsi qu'en composés volatiles, en caroténoïdes et en composés phénoliques (Bindon et al., 2007; Castellarin et al., 2007a,b; Deluc et al., 2009; Griesser et al., 2015; Hochberg et al., 2015; Savoï et al., 2016; Marfil et al., 2019). Toutefois ces tendances ont généralement été mises en évidence sur un nombre limité de cépages et leur variabilité génétique demeure peu explorée.

Dans ce contexte, mon projet de recherche vise à développer et utiliser des ap-

1. stade correspondant à une augmentation rapide du sucre dans les baies accompagnée généralement d'un changement de couleur et/ou d'un ramollissement.

proches de phénotypage à haut-débit, notamment basées sur des endophénotypes (phénotypes moléculaires intermédiaires entre génome et phénotypes mesurés à l'échelle des organismes), pour caractériser la variabilité génétique de la réponse de la vigne à des contraintes exercées par le changement climatique comme le stress hydrique. Ces approches vont jusqu'à l'étude des déterminismes génétiques des caractères (phénotypes à l'échelle de l'organisme et endophénotypes), et l'objectif final est de fournir des outils et méthodologies pour les programmes d'amélioration génétique de la vigne. Pour mener à bien ces travaux, je m'appuie sur des expertises développées dans le cadre de mes travaux de recherche précédents ainsi que sur des collaborations notamment avec des collègues (eco)-physiologistes. Les compétences que je vais notamment pouvoir mobiliser sont : le phénotypage à haut-débit par NIRS, la génétique quantitative, et plus spécifiquement génétique d'association et prédiction génomique, la prédiction phénotypique et l'intégration multi-omiques.

3.2 Architecture génétique de caractères d'intérêt en réponse à la contrainte hydrique

3.2.1 Matériel d'étude : un panel de diversité génétique

Pour mettre en œuvre cette partie, je m'appuie principalement sur un panel développé au sein de l'équipe DAAV pour représenter la diversité génétique de la vigne cultivée (Nicolas et al., 2016). On peut noter qu'il s'agit du même panel que celui évoqué dans la partie bilan pour les travaux de prédiction phénotypique chez la vigne effectués dans le cadre de la thèse de Charlotte Brault (chapitre 2.4.2). Ce panel comprends 279 variétés qui se structurent en trois grandes sous-populations représentées de façon équivalente : 93 variétés à raisins de cuve originaires de l'Ouest de l'Europe (Wine West, WW), 93 variétés à raisins de cuve originaires de l'Est de l'Europe (Wine East, WE), et 93 variétés à raisins de table originaires de l'Est de l'Europe (Table East, TE).

Ce panel a été génotypé par puce ainsi que par séquençage après réduction de la complexité du génome, fournissant un peu plus de 90 000 SNPs (Flutre et al., 2020). Par ailleurs, un sous-échantillon de 60 variétés a fait l'objet d'un re-séquençage complet dans le cadre d'un projet plus large, et je prévois de tester une augmentation de la densité de génotypage sur les 219 variétés restantes par imputation. Pour cela, une stratégie par validation croisée sera d'abord mise en œuvre sur les 60 variétés reséquencées avec potentiellement plusieurs outils, afin de déterminer la qualité d'imputation de chacune des positions. Dans un deuxième temps, l'imputation sera déployée sur l'ensemble des variétés du panel avec l'outil ayant permis d'obtenir la meilleure qualité d'imputation globale en validation croisée. L'ensemble des positions sera imputée, mais les positions seront par ailleurs annotées avec leur valeur de précision d'imputation, telle que déterminée dans la validation croisée. Cela permettra par la suite de sélectionner des sous-ensembles de SNPs selon ce critère, que l'on pourrait faire varier en fonction des objectifs scientifiques. Ce type de straté-

gie a déjà été utilisée avec succès chez le peuplier (Pégard et al., 2020) ou chez *Arabidopsis thaliana* (Arouisse et al., 2020).

Le panel de diversité a par ailleurs déjà fait l'objet d'un suivi phénotypique au vignoble pour un ensemble de caractères d'intérêt, décrivant la phénologie, la vigueur, le rendement, la morphologie de la grappe et la composition métabolique du raisin (Pinasseau et al., 2017). De plus, une première étude de génétique d'association à partir de ces données a conduit à l'identification de nombreux QTLs pour ces caractères (Flutre et al., 2020). On peut noter que ce dispositif avait été soumis au cours de deux années successives à une modalité d'irrigation différenciée avec trois blocs irrigués et deux blocs non irrigués, afin d'étudier la variabilité génétique de la composition en polyphénols du raisin en réponse au stress hydrique. Si ces travaux ont montré des réponses différentielles au stress hydrique selon les familles de polyphénols et les génotypes (Pinasseau et al., 2017), les premiers travaux de génétique d'association se sont plutôt focalisés sur les régions génomiques stables indépendamment du stress, plutôt que sur celles qui pourraient être associées à la réponse des génotypes à l'irrigation différenciée. Par ailleurs, les autres caractères d'intérêt avaient été évalués au cours d'autres saisons de végétation, ce qui ne permet pas d'étudier l'impact du stress d'une part sur les autres caractères d'intérêt comme le rendement, la vigueur ou la phénologie, et d'autre part sur les relations entre fonctionnements végétatif et reproducteur. Le projet de recherche G2WAS ambitionne notamment de répondre à ces questions.

3.2.2 Le projet G2WAS

Plus spécifiquement, le projet G2WAS vise à élucider les bases génétiques des caractères impliqués dans la réponse de la vigne à la contrainte hydrique avec un focus particulier d'une part sur les relations entre fonctionnements végétatif et reproducteur au sein de la même saison, et d'autre part sur les effets inter-annuels du stress (sur la récolte de l'année suivant la contrainte hydrique). Pour cela, le panel de diversité va être suivi en 2022 en conditions semi-contrôlées au sein de la plateforme PHENOARCH de l'UMR LEPSE, avec au moins une modalité témoin et une modalité stressée. Dans ce large projet collaboratif, je serai spécifiquement impliqué dans l'étude de caractères de fonctionnement végétatif et du métabolisme des baies. Je co-coordonne par ailleurs, avec Tristan Mary-Huard (UMR GQE et MIA-Paris) le work-package dédié aux analyses statistiques des données produites. Pour la génétique d'association, nous utiliserons notamment des modèles multivariés, comme le modèle MTMM (Korte et al., 2012), pour identifier des polymorphismes spécifiquement impliqués dans la réponse au stress.

3.2.3 Fonctionnement végétatif

Les travaux sur le fonctionnement végétatif sont effectués en collaboration avec l'équipe ETAP de l'UMR LEPSE et plus spécifiquement Aude Coupel-Ledru.

Pour pouvoir étudier la variabilité de caractères de fonctionnement végétatif au

sein de larges populations, telles que celles typiquement étudiées dans les études de variabilité génétique et les programmes d'amélioration, il est crucial de pouvoir disposer d'outils de phénotypage à haut-débit. La réflectance des tissus végétaux dans le visible et le proche infrarouge est utilisée depuis de nombreuses années pour calculer des indices de végétation, comme par exemple le NDVI (Normalized Vegetation Index) ou le PRI (Physiological Reflectance Index), qui témoignent de l'activité photosynthétique du couvert (Tucker, 1979; Gamon et al., 1992). Plus récemment, des travaux utilisant des spectres de réflectance ou d'absorbance, pouvant aller du visible jusqu'au proche infrarouge, montrent qu'il est possible de développer des calibrations pour de nombreux caractères fonctionnels de la même manière que l'on réalise des calibrations pour des propriétés physico-chimiques (Grzybowski et al., 2021).

Pour tester cette approche sur vigne, nous avons initié une expérience de calibration basée sur des feuilles échantillonnées dans deux dispositifs complémentaires : un plan de croisement demi-diallèle au vignoble et le panel de diversité en pots. Le plan de croisement est composé de 10 familles issues de 5 cépages : Cabernet Sauvignon, Grenache, Syrah, Pinot Noir et Terret Noir. Chaque famille inclus environ 60 génotypes implantés au domaine du Chapitre (Villeneuve-lès-Maguelone, 34) à raison de 2 répétitions de parcelles unitaires de 2 souches réparties dans 2 blocs complets randomisés. On peut noter que ce dispositif a aussi déjà été évoqué dans la partie bilan pour les travaux de prédiction phénomique chez la vigne effectués dans le cadre de la thèse de Charlotte Brault (chapitre 2.4.2). Le panel de diversité correspond au panel comprenant 279 génotypes qui sont conduits en pots sur le campus de La Gaillarde de Montpellier SupAgro. Il s'agit des mêmes plantes qui seront étudiées au sein de la plateforme PHENOARCH dans le cadre du projet G2WAS.

Ce travail de calibration a fait l'objet du stage de Master 1 de Virgilio Freitas que j'ai co-encadré avec Aude Coupel-Ledru en 2021. En pratique, nous avons échantillonné environ 220 feuilles des deux dispositifs, sur lesquelles nous avons mesuré *in situ* :

- les échanges gazeux et la fluorescence chlorophyllienne au moyen d'un dispositif de référence à débit relativement bas (Li-6800, LI-COR Biosciences Inc., Lincoln, NE) ;
- la conductance stomatique et la fluorescence chlorophyllienne au moyen d'un fluoromètre/poromètre récemment développé et à débit élevé (Li-600, LI-COR Biosciences Inc.) ;
- la réflectance dans le proche infrarouge au moyen de deux spectromètres micro-portables aux gammes de longueur d'ondes complémentaires (Micro-NIR OnSite-W, Viavi et NeoSpectra Scanner, Sci-Ware) ;
- la teneur en chlorophylle au moyen d'un chlorophylle-mètre portable basé sur l'absorbance dans une gamme de longueur d'onde restreinte entre le visible et le proche infrarouge (SPAD, Konica Minolta).

Nous avons également collectés sur chacune des feuilles des disques foliaires pour évaluer la masse surfacique (Specific Leaf Area, SLA) et collecter des spectres au

laboratoire sur une gamme spectrale plus large incluant visible et proche infrarouge. Ces spectres au laboratoire ont été collectés en collaboration avec Martin Ecartot (équipe GE²pop, UMR AGAP Institut) et donc avec le même spectromètre (LabSpec, ASD) que celui utilisé pour les prédictions phénomiques dans le cadre de la thèse de Charlotte Brault. Par ailleurs, pour le dispositif en pots seulement, un certain nombre d’analyse supplémentaires ont pu être effectuées, comme par exemple la détermination du potentiel hydrique d’une feuille voisine.

Des premières calibrations NIRS ont été établies par régression PLS pour les caractères de référence. Ces calibrations permettent d’atteindre des précisions de prédiction relativement satisfaisantes, notamment pour la masse surfacique ou l’assimilation nette du carbone, ce qui confirme les potentialités de la NIRS comme outil de phénotypage à haut-débit de caractères fonctionnels chez la vigne (FIGURE 19). On peut par ailleurs noter des différences entre les modalités d’acquisition des spectres pour certains caractères, comme l’assimilation nette du carbone qui est mieux prédite avec les spectres collectés sur disques séchés au laboratoire ou le potentiel hydrique foliaire qui à l’inverse est mieux prédit avec les spectres collectés *in situ*.

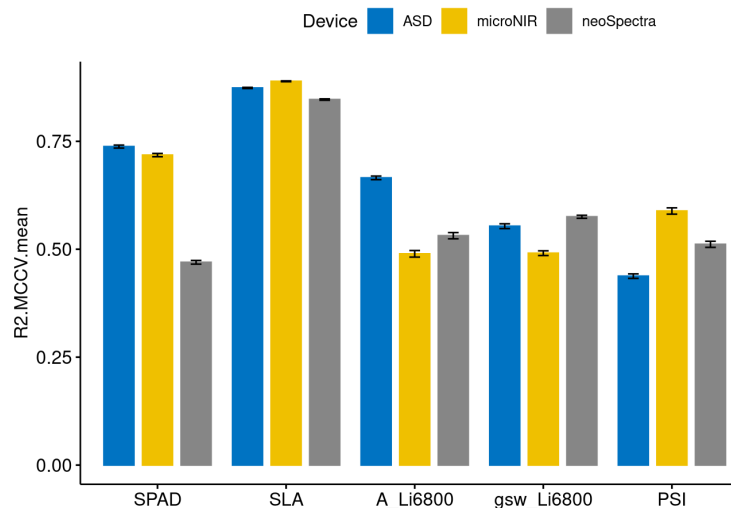


FIGURE 19 – Résultats des calibrations NIRS pour des caractères fonctionnels chez la vigne.

Les couleurs correspondent aux modalités d’acquisition des spectres, ASD étant une acquisition au laboratoire sur disques foliaires séchés avec un spectromètre qui couvre le visible et le proche infrarouge, tandis que microNIR et neoSpectra sont des acquisitions *in situ* avec des spectromètres micro-portables qui couvrent des gammes de longueurs d’onde relativement restreintes et complémentaires dans le proche infra-rouge. Les caractères fonctionnels prédits sont le SPAD, un proxy de la teneur en chlorophylle, la masse surfacique des feuilles (SLA), l’assimilation nette du carbone (A_Li6800), la conductance stomatique (gsw_Li6800) et le potentiel hydrique foliaire (PSI).

Ces calibrations vont être étendues à une plus large gamme de caractères, notamment biochimiques (teneurs en sucres, azote et carbone) qui sont en cours d’ac-

quisition sur des disques collectés sur les mêmes feuilles. Des analyses métabolomiques sont également prévues afin d'étendre la gamme des composés étudiés et ainsi aller plus loin dans l'interprétation de ce qui relie les caractères fonctionnels aux spectres. Ces analyses devraient également permettre d'identifier des marqueurs métaboliques du fonctionnement des plantes. Elles seront effectuées en collaboration avec le plateau d'analyse du métabolisme secondaire de la vigne de l'UMR SVQV à Colmar. Enfin, les calibrations NIRS seront également étendues à ces données métabolomiques.

Au-delà de ces calibrations, nous avons collecté, en 2021, des spectres dans le proche infrarouge sur l'ensemble des géotypes des 2 dispositifs, de façon à prédire puis analyser la variabilité génétique des caractères fonctionnels. Par ailleurs, pour ce qui est des spectres collectés sur disques foliaires au laboratoire, d'autres analyses avaient été effectuées en 2020 et 2021 sur les dispositifs au vignoble du plan de croisement demi-diallèle et du panel de diversité afin d'évaluer les performances de la prédiction phénomique dans le cadre de la thèse de Charlotte Brault (chapitre 2.4.2). Ces spectres pourront être valorisés par des prédictions des caractères fonctionnels pour étendre les analyses de leur variabilité génétique. En effet, l'ensemble des données prédites seront utilisées pour estimer l'héritabilité des caractères fonctionnels dans les différents dispositifs et entreprendre des détectations de QTL, soit par cartographie génétique au sein du plan de croisement demi-diallèle, soit par génétique association dans le panel de diversité.

Par ailleurs, les calibrations développées en 2021 seront mises à jour en 2022 au cours de l'expérimentation prévue dans le cadre de G2WAS au sein de la plateforme PHENOARCH. Cela permettra de tester la robustesse des modèles notamment lorsque les plantes sont soumises à un stress hydrique. Des spectres seront également collectés sur des feuilles de l'ensemble des plantes du dispositif en vue de pouvoir utiliser les calibrations mises à jour pour prédire les caractères fonctionnels et biochimiques des feuilles et ainsi évaluer leur variabilité au sein de la population et en réponse à la contrainte hydrique.

Pour mener à bien ces travaux, nous allons co-encadrer, avec Aude Coupel-Ledru, un stage de fin d'étude, qui pourrait se poursuivre par une thèse sur ce sujet. En plus des calibrations et des analyses de la variabilité et du déterminisme génétique des caractères fonctionnels en réponse au stress hydrique, cette thèse pourrait inclure des validations au vignoble ou du screening dans le cadre des programmes d'amélioration génétique de l'arc méditerranéen, pour lesquels l'adaptation au stress hydrique est un enjeu prioritaire. Nous prévoyons notamment de déposer un projet en lien avec l'interprofession dans les mois à venir pour soutenir spécifiquement ces travaux.

3.2.4 Métabolisme de la baie

L'autre volet de perspectives, sur l'architecture génétique de caractères d'intérêt en réponse à la contrainte hydrique, concerne le métabolisme de la baie. Ces travaux sont effectués en collaboration avec Charles Romieu de l'équipe DAAV. Ils font notamment l'objet de la thèse de Flora Tavernier que nous co-encadrons et qui

vient de débuter en Novembre 2021.

Les travaux précédemment réalisés dans le dispositif au vignoble ont montré une certaine variabilité de la composition en polyphénols au sein du panel et que cette variabilité était modulée par la contrainte hydrique (Pinasseau et al., 2017). D'autres travaux ont montré que le stress hydrique au cours de la maturation du raisin pouvait affecter sa composition en métabolite primaires et secondaires, mais ces travaux ont souvent été limités à quelques génotypes (Bindon et al., 2007; Castellarin et al., 2007a,b; Deluc et al., 2009; Griesser et al., 2015; Hochberg et al., 2015; Savoï et al., 2016; Marfil et al., 2019). Dans ce contexte, l'expérimentation du projet G2WAS en conditions semi-contrôlées dans PHENOARCH offrait une opportunité pour pouvoir étudier la variabilité de la réponse métabolique du raisin à une contrainte hydrique et au sein d'un large panel de diversité.

L'échantillonnage des baies à analyser sera une des clés de la réussite de ce projet. En effet, les grappes de raisin se caractérisent par une grande variabilité phénologique entre baies qui est typiquement visible au stade véraison sur les variétés à raisins rouges. Cette variabilité ne suit pas de patron particulier sur la grappe, et doit être prise en compte pour éviter de confondre les différences de composition observées avec des différences liées au stade de développement (FIGURE 20).

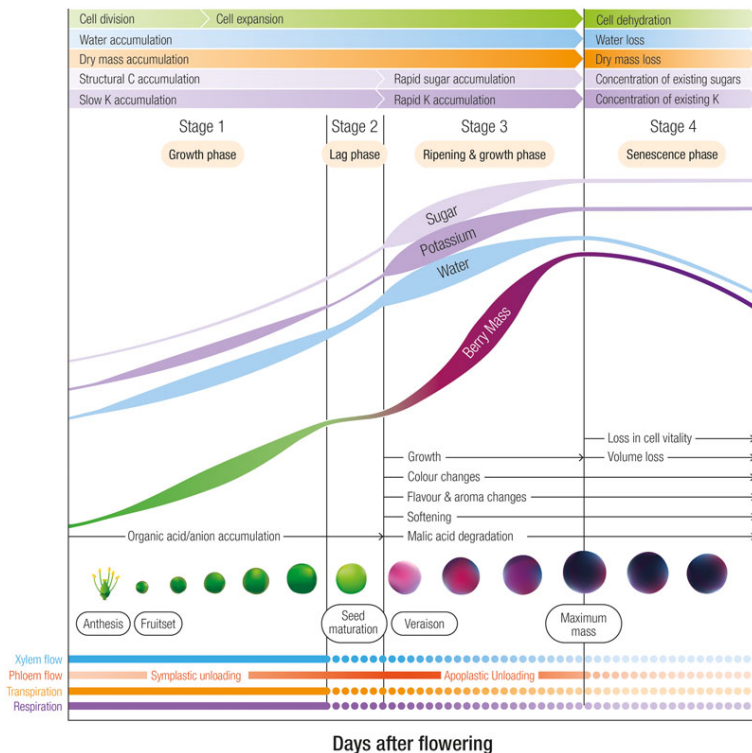


FIGURE 20 – Stades de développement d'une baie. Représentation schématisée des stades de développement d'une baie avec les principaux changements associés aux différentes phases. D'après (Rogiers et al., 2017).

Dans le projet G2WAS la contrainte hydrique sera appliquée à partir de la vé-

raison et devrait ainsi affecter la période de maturation des baies. Nous visons la fin de la période de maturation pour effectuer l'échantillonnage qui permettra d'évaluer l'impact de la contrainte sur la composition en métabolites du raisin. A l'échelle de la baie, ce stade peut-être déterminé soit en suivant l'évolution du volume par analyse d'image, soit en suivant la dynamique d'évolution de certains composés comme le sucre ou l'eau. Dans ce dernier cas, nous testerons la NIRS comme méthode non destructive de suivi des principaux changements qui ont lieu au cours de la maturation d'une baie. La combinaison de ces deux approches d'imagerie et de spectroscopie devrait permettre de bien maîtriser et caractériser le stade de développement des baies prélevées. Ces dernières seront ensuite broyées et lyophilisées avant dosage des métabolites primaires (sucres et acides) et secondaires par chromatographie liquide couplée à un spectromètre de masse (LC-MS) pour l'analyse des composés non volatils. Cette technologie permet de doser une très large gamme de composés dont des acides aminés, des flavonoïdes, des stilbènes et des composés phénoliques. Comme dans le cas des feuilles, les analyses métabolomiques seront réalisées en collaboration avec le plateau d'analyse du métabolisme secondaire de la vigne de l'UMR SVQV à Colmar.

Les questions suivantes pourront être adressées avec ce large jeu de données métabolomiques collectées sur des plantes témoins et stressées de l'ensemble des génotypes du panel :

- Quels métabolites sont impliqués dans la signalétique de l'adaptation au stress hydrique ?
- Quel est le niveau de variation et de co-variation génétique de ces métabolites en réponse au stress ?
- Quels sont les régions génomiques et les gènes candidats associés à cette variabilité ?

Pour répondre à ces questions, des analyses multivariées seront mises en œuvre ainsi que des analyses de génétique quantitative. Comme dans le cas des traits de fonctionnement foliaire, nous utiliserons le modèle linéaire mixte multivarié pour partitionner la variance observée pour chaque métabolite en variance génétique, variance d'interaction entre génétique et contrainte hydrique et variance résiduelle. Un tri sur le ratio entre variance d'interaction génétique \times stress et variance génétique permettra de mettre en évidence les métabolites les plus impliqués dans la réponse au stress. Des analyses d'association avec le modèle MTMM seront également mises en œuvre afin d'identifier spécifiquement les polymorphismes sous-jacents à l'interaction génétique \times stress (Korte et al., 2012; Albert et al., 2016). Sous l'hypothèse que l'analyse multivariée permettra de mettre en évidence des groupes de métabolites corrélés au sein du panel, des analyses d'association par groupe seront aussi mises en œuvre pour identifier spécifiquement des loci "hubs", c'est à dire associés à de nombreux métabolites, de la même manière que cela peut-être observé dans les analyses eQTLs. Ces loci pourraient être intéressants en sélection, car à effet potentiellement fort sur des phénotypes plus complexes, du fait qu'ils contrôlent la variation d'un grand nombre de métabolites.

Au-delà des approches de génétique quantitative, il sera possible de tester des calibrations NIRS pour chacun des métabolites avec les spectres collectés sur les baies avant récoltes, mais aussi des spectres collectés au laboratoire sur les mêmes échantillons que ceux analysés par LC-MS. Ces calibrations, sous l'hypothèse qu'elles soient suffisamment précises, pourraient permettre de s'affranchir au moins en partie des analyses métabolomiques relativement coûteuses pour quantifier les molécules qui varient lors d'un stress hydrique et qui pourraient être des marqueurs d'adaptation des variétés à cette contrainte. Dans ce cas de figure, le haut-débit et le faible coût de la méthode NIRS pourrait permettre de les déployer dans les programmes d'amélioration.

3.2.5 A plus long terme : validations au vignoble et analyses de diversité entre génotypes extrêmes

Néanmoins, il ne faut pas oublier que ces expérimentations se feront sur des plantes en pots dans des conditions semi-contrôlées. Avant de pouvoir en utiliser les principaux résultats dans les programmes de sélection variétale, il faudra passer par des étapes de validation au vignoble.

Différents dispositifs pourront être utilisés dans cet objectif. Tout d'abord, le panel de diversité précédemment décrit a été implanté en 2021 au domaine de Pech-Rouge (Gruissan, 11), à raison de 4 blocs complets randomisés. Ce dispositif, lorsqu'il aura atteint sa maturité dans quelques années, fera l'objet d'une irrigation différenciée avec 2 blocs irrigués et 2 blocs non irrigués. A plus long terme, ce même dispositif devrait être déployé sur d'autres sites en France mais aussi à l'International, ouvrant pas mal de perspectives sur l'étude des IGE. Par ailleurs, dans le cadre du projet SelGenVit (ANR-19-ECOM-0006) un autre panel est en cours d'implantation sur trois sites constituant un gradient latitudinal à l'échelle de la France : Montpellier, Villefranche-sur-Saône, et Colmar. Ce panel de 200 variétés a été constitué pour être représentatif de la diversité utilisée dans les programmes de sélection de la vigne et ainsi pour servir de population d'entraînement pour des modèles de prédiction génomique ou phénotypique. Enfin, d'autres dispositifs des programmes régionaux de l'arc méditerranéen (Provence, Côtes du Rhône, Languedoc) pourraient être utilisés pour valider les différents marqueurs de la réponse des génotypes de vigne au stress hydrique que ce soit à l'échelle du fonctionnement foliaire ou de la composition du raisin.

Des approches complémentaires de génomique des populations pourraient aussi être mises en œuvre pour progresser sur l'identification de régions génomiques d'intérêt pour l'adaptation de la vigne aux contraintes induites par le changement climatique. En effet, une core-collection, appelé C4, a spécifiquement été définie dans l'équipe pour conduire des études sur le changement climatique (Boursiquot et al., 2018). Ce panel a été constitué à partir de données issues de la collection de référence (Conservatoire des Ressources Biologiques-Vigne de Vassal-Montpellier). Il inclut notamment des génotypes aux performances extrêmes pour les caractères les plus impactés par le changement climatique, comme par exemple la phénologie ou

la tolérance aux stress abiotiques. Des études de différenciation génomique entre groupes de génotypes extrêmes pourraient permettre d'identifier des loci impliqués dans l'adaptation, d'autant que le génome d'un certain nombre de variétés de ce panel a récemment fait l'objet d'un re-séquençage complet. Cette approche pourrait être complétée par des analyses transcriptomiques ou métabolomiques qui rechercheraient des gènes différentiellement exprimés ou des métabolites variables entre ces mêmes groupes de génotypes extrêmes. Cette approche serait ainsi assez complémentaire de celle déployée sur le panel de diversité.

3.3 Outils d'aide à la sélection

Cette partie concerne mes perspectives en lien avec les programmes d'amélioration génétique de la vigne. Ces travaux s'intègrent notamment dans le projet de l'UMT (Unité Mixte Technologique) Géno-Vigne[®]3 entre INRAE, l'IFV (Institut Français de la Vigne et du Vin) et Montpellier SupAgro. Ils se font donc principalement en collaboration avec Loïc Le Cunff (ingénieur généticien, IFV). Cela concerne notamment l'utilisation de données génomiques et phénotypiques pour la prédiction de caractères complexes.

3.3.1 Vers l'utilisation des prédictions génomique et phénotypique dans les programmes d'amélioration génétique de la vigne

En ce qui concerne la prédiction génomique, si cette méthodologie, proposée il y a une vingtaine d'années dans le domaine de la génétique animale, est désormais appliquée en routine chez de nombreuses espèces d'élevage et de grandes cultures. Chez les espèces végétales pérennes, cette méthodologie fait encore l'objet de travaux de recherche pour déterminer les conditions optimales de son application en sélection. Chez la vigne en particulier, des premiers travaux effectués au sein de l'équipe DAAV, basés sur des simulations, avaient montré son intérêt potentiel (**Fodor et al., 2014**). D'autres travaux, sur données réelles, ont montré qu'il était possible de prédire de nombreux caractères d'intérêt avec des précisions de prédiction moyennes à élevées selon le caractère considéré (Migicovsky et al., 2017; Flutre et al., 2020). Toutefois ces travaux étaient basés sur des validations croisées dans des panels de diversité et l'application de la prédiction génomique dans un contexte inter-populationnel plus proche de la réalité des programmes d'amélioration restait à tester afin de promouvoir son utilisation. Ce travail a spécifiquement fait l'objet d'un chapitre de la thèse de Charlotte Brault.

En se basant sur des données disponibles, notamment sur le panel de diversité et le plan de croisement demi-diallèle précédemment décrits, Charlotte a pu spécifiquement tester un scénario avec entraînement des modèles prédictifs dans le panel de diversité et validation dans les croisements biparentaux du demi-diallèle (**Brault et al., soumis**). Ce scénario se rapproche de ce qui pourrait être utilisé pour l'amélioration génétique d'une espèce pérenne végétale, avec une population d'entraînement qui capture de la diversité génétique et des populations candidates

à la sélection issue de croisements. Outre l'intérêt de ce scénario, le fait de faire la validation dans un plan de croisement a permis de décomposer les précisions de prédiction en valeur moyenne du croisement et en valeur génétique individuelle des descendants au sein de chaque croisement (échantillonnage mendélien). Cette décomposition est particulièrement intéressante puisqu'elle permet de se projeter dans une utilisation de la prédiction génomique à deux niveaux du programme de sélection : d'abord pour définir *in silico* les croisements à réaliser sur la base de la prédiction de leur valeur moyenne, puis dans une utilisation plus classique, une fois les croisements réalisés, pour prédire les valeurs génétiques des individus au sein des croisements. Ces travaux montrent des précisions de prédiction génomique qui varient selon le caractère mais qui sont globalement encourageantes tant pour la moyenne des croisements et que pour les performances des individus au sein des croisements. La prédiction de la moyenne a par ailleurs été complétée par une prédiction de la variance, évaluée par validation croisée dans un scénario intra-populationnel, qui est aussi tout à fait encourageante dans l'objectif de prédiction *in silico* des croisements à réaliser.

Bien que ce travail représente une étape importante en vue de déployer la prédiction génomique dans les programmes d'amélioration de la vigne, certaines limites demeurent, notamment en lien avec le panel d'entraînement utilisé. En effet, comme précédemment mentionné, ce panel a été constitué pour capturer la diversité génétique de la vigne cultivée, et de fait il n'inclut pas de résistance aux principales maladies de la vigne. Ces résistances sont introgressées depuis le compartiment sauvage via des géniteurs hybrides et la précision des prédictions dans ce contexte particulier reste à tester. C'est pour répondre spécifiquement à cette limite qu'un nouveau panel a été défini dans le cadre du projet SelGenVit (ANR-19-ECOM-0006), coordonné par Komlan Avia (UMR SVQV, Colmar) et dans lequel je suis en charge du workpackage dédié à la production de données. Le panel inclus environ 200 génotypes qui sont les parents des croisements actuels et futurs des programmes d'amélioration en France. Il s'agit notamment de cépages emblématiques des grandes régions viticoles françaises, de variétés hybrides résistantes au mildiou et l'oïdium, de cépages identifiés comme présentant des caractères favorables dans le contexte du changement climatique. Comme précédemment mentionné, ce panel est en cours d'implantation au sein de 3 grandes régions constituant un gradient latitudinal : (1) Alsace avec 2 sites sur Colmar ; (2) Beaujolais au domaine de la SICAREX à proximité de Villefranche-sur-Saône ; (3) Languedoc au domaine Maspique à proximité de Montpellier. Dans le cadre du projet SelGenVit, ce panel fera l'objet d'un génotypage, d'analyses NIRS et métabolomiques sur feuilles, ainsi que de premières évaluations phénotypiques pour la phénologie notamment. Pour ce qui est du génotypage, une stratégie basée sur du re-séquençage et de l'imputation est en cours de réflexion puisqu'un certain nombre d'individus du panel ont récemment été re-séquencés dans le cadre d'un projet plus large. En ce qui concerne la NIRS, nous poursuivrons les investigations, initiées lors de la thèse de Charlotte Brault, sur la prédiction phénotypique. Il s'agira notamment de tester des prédictions basées sur des acquisitions *in situ* avec des spectromètres micro-portables et sur des plantes

jeunes en vue notamment de prédire les caractères de production au stade mature.

Un autre aspect concerne l'intégration multi-omiques pour la prédiction. Les premiers essais sur vigne dans le cadre de la thèse de Charlotte Brault sont un peu décevants, puisque les modèles combinant SNPs et NIRS font généralement aussi bien que les meilleurs modèles basés seulement sur les SNPs ou les spectres NIRS. Cette tendance avait déjà été observée dans l'article original sur la prédiction phénomique (**Rincent et al., 2018**), et les travaux effectués dans le cadre de la thèse d'Abdou Rahmane Wade sur l'intégration de SNPs et transcrits offrent un certain éclairage sur les conditions favorables à l'intégration multi-omiques (**Wade et al., soumis**). De façon similaire, des analyses permettant d'évaluer la redondance entre matrices de prédicteurs, afin notamment de définir une stratégie qui permette de la prendre en compte explicitement dans le modèle, pourraient être développées dans le contexte de la prédiction phénomique. Les données métabolomiques offrent une certaine opportunité dans ce sens, puisque contrairement aux spectres, il s'agit là directement de molécules pour lesquelles des détections de QTLs pourront être effectuées pour quantifier les liens entre prédicteurs génomiques et métabolomiques voire phénomiques.

Par ailleurs, le fait d'avoir un dispositif multi-sites permettra spécifiquement d'adresser la question de la précision des prédictions génomique et phénomique dans un contexte multi-environnemental. Pour cela, il sera intéressant de compléter les données existantes par des données environnementales pour les utiliser comme co-variables dans les modèles de prédiction génomique (Heslot et al., 2014; Jarquín et al., 2014). Les données NIRS et métabolomiques pourraient aussi jouer un rôle particulier dans la prise en compte des IGE au-delà de la simple prédiction phénomique. En effet, contrairement aux marqueurs moléculaires, les métabolites et par conséquent les spectres dans le proche infrarouge peuvent témoigner de la réponse des plantes aux conditions environnementales, comme peuvent en témoigner les indices de végétation (Tucker, 1979; Gamon et al., 1992) ou les calibrations NIRS pour des caractères fonctionnels (Grzybowski et al., 2021). Si cette instabilité peut sembler au premier abord néfaste pour le succès de la prédiction phénomique, elle souligne aussi que les spectres sont capables de capturer les réponses génétiques à un environnement donné, ce qui ouvre de nouvelles perspectives d'application pour la prédiction phénomique de l'IGE. On peut d'ailleurs noter ici que dans l'article original sur la prédiction phénomique, nous avons estimé une composante d'IGE le long du spectre qui pouvait atteindre jusqu'à 40% de la variance totale pour certaines longueurs d'ondes (**Rincent et al., 2018**).

Ainsi pour la prédiction de l'IGE, les spectres ou métabolites collectés dans chaque environnement pourraient par exemple être utilisés pour estimer des matrices de covariance entre génotypes spécifiques à chaque environnement (matrices de similarités phénomiques). Krause et al. (2019) et Lane et al. (2020) ont notamment utilisé cette approche à partir de données hyperspectrales, pour prédire le rendement en grain dans des essais multi-environnementaux, chez le blé et le maïs respectivement. Ils ont montré que l'utilisation de données spectrales améliore la précision de prédiction des IGE par rapport aux modèles qui utilisent seulement

les marqueurs moléculaires ou le pedigree. Une autre possibilité serait d'utiliser les spectres pour estimer des similarités entre les différents environnements, comme proposé par Heslot et al. (2014) et Jarquín et al. (2014) avec des covariables environnementales. Ces différentes applications pourraient donc permettre de faire des prédictions dans de nouveaux environnements, dans lesquels seuls les spectres seraient collectés, en estimant des matrices de similarité phénotypique entre génotypes ou entre environnements. Ce type de configuration pourra notamment être testé chez la vigne dans un scénario inter-populationnel, comme dans le cadre de la thèse de Charlotte Brault, en utilisant pour l'entraînement des modèles, les dispositifs du projet SelGenVit, et pour la validation, les dispositifs en stade intermédiaire de l'interprofession. Ces derniers dispositifs correspondent en particulier aux croisements d'intégration qui font l'objet de ce qui suit.

3.3.2 Les cas particulier des croisements d'intégration

Le deuxième volet de perspectives en lien avec les programmes d'amélioration concerne les travaux mis en œuvre dans le cadre du projet OASIs (CASDAR) qui a débuté en 2021 et que je coordonne. Ce projet cible spécifiquement le développement d'outils d'aide à la sélection pour les croisements dits d'intégration qui ont été initiés en 2014 par INRAE et l'IFV entre des cépages emblématiques des grandes régions viticoles françaises et des variétés résistantes aux principales maladies (FIGURE 21).

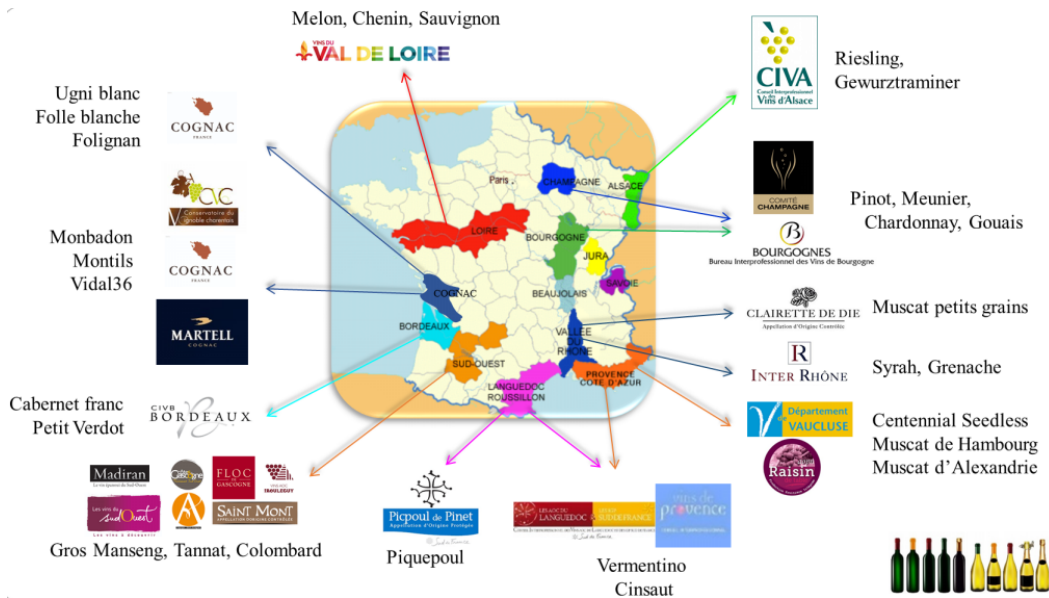


FIGURE 21 – Dynamique du programme d'amélioration de la vigne en France. Carte représentant la dynamique avec les Interprofessions, Syndicats et Entreprises impliqués dans les croisements d'intégration ainsi que les cépages emblématiques choisis comme un des géniteurs.

L'objectif de ce programme est de sélectionner des descendants résistants qui présentent des caractéristiques viti-œnologiques proches de celles de leur parent cépage emblématique. Si les premières sélections des descendants résistants sont relativement rapides car elles se basent sur des marqueurs moléculaires pour les résistances, les évaluations phénotypiques au vignoble sont généralement longues (environ 6 ans) et coûteuses car elles doivent être réalisées sur 5 à 10 répétitions par génotype (FIGURE 17). Le projet OASIs ambitionne de développer des outils d'aide à la sélection pour accélérer le processus de création variétale dans le cadre des croisements d'intégration.

Plus spécifiquement, j'ai proposé dans ce projet de tester l'intérêt du phénomène, évalué par NIRS sur différents organes, et des marqueurs moléculaires, pour trier les descendants résistants des croisements selon leur proximité à leur parent cépage emblématique. Pour cela, nous utiliserons spécifiquement les dispositifs en stade intermédiaire de 3 interprofessions partenaires du projet : le Comité Interprofessionnel du Vin de Champagne, l'Institut Rhodanien (Côtes-du-Rhône), et le Bureau National Interprofessionnel du Cognac. En ce qui concerne les marqueurs, nous passerons par une étape préalable d'identification de traces de sélection le long du génome, sur la base des sélections réalisées dans les programmes d'amélioration par les 3 interprofessions. Nous pourrions aussi bénéficier des données phénotypiques collectées sur ce matériel pour effectuer des détectations de QTL et croiser cette information sur le déterminisme génétique des caractères avec les traces de sélection.

Dans la première partie du projet, nous avons proposé de faire des développements méthodologiques en nous basant sur le dispositif demi-diallèle précédemment décrit. Ces développements concernent à la fois la collecte des spectres (tissu et modalité d'acquisition au laboratoire ou *in situ*) et la détection de traces de sélection. Pour cela, nous nous appuyons sur un certain nombre de données phénotypiques, génotypiques et spectrales déjà disponibles (Tello et al., 2019, **Brault et al., soumis**). Ces données sont en cours de complémentation avec des spectres collectés *in situ* sur feuilles et grappes en 2021. Dans un deuxième temps, nous effectuerons des analyses NIRS au laboratoire sur des échantillons de feuilles, bois et jus, provenant des dispositifs en stade intermédiaire et des analyses NIRS *in situ* sur feuilles et grappes. Nous réaliserons également le génotypage des individus pour les analyses basées sur les marqueurs moléculaires. Enfin ces données seront analysées pour inférer des distances entre les descendants résistants et leur parents, et déterminer ceux qui sont les plus proches des cépages emblématiques de chacune des 3 régions. Au-delà du projet, ces travaux pourront être déployés dans d'autres programmes régionaux. Par ailleurs, en lien avec ma première partie de perspectives, les spectres NIRS collectés dans les programmes régionaux pourront être utilisés pour prédire le fonctionnement des individus en cours sélection dans un contexte hydrique limitant et ainsi inclure dans les critères de sélection des caractères d'intérêt pour l'adaptation au changement climatique.

3.4 Conclusions sur la partie perspectives

Cette partie perspective met en évidence des activités à l'interface entre déterminisme génétique et prédictions génomiques et phénotypiques dans le contexte de l'adaptation au changement climatique chez la vigne. Deux projets de thèse (une en cours et une en prévision), associées au projet G2WAS, permettront notamment de mener à bien les activités concernant les déterminismes génétiques des caractères liés au fonctionnement foliaire et au métabolisme de la baie en réponse à une contrainte hydrique. En ce qui concerne les activités plus appliquées, elles sont notamment supportées par deux projets en cours, SelGenVit et OASIs, avec des dispositifs complémentaires. En effet, dans SelGenVit il s'agit du panel pour l'entraînement des modèles de prédiction génomique, tandis que dans OASIs il s'agit des variétés en cours de sélection. Le panel d'entraînement étant en cours d'implantation, nous commençons à réfléchir à des dépôts de projets pour son phénotypage dans les différents sites et ainsi commencer à développer et tester les prédictions génomiques et phénotypiques dans un contexte multi-environnemental. A moyen terme, cela pourrait faire l'objet d'une thèse en lien avec l'interprofession, et qui serait spécifiquement axée sur la prise en compte des IGE dans les modèles prédictifs.

L'ensemble de ces travaux devrait permettre de compléter les données actuellement disponibles par des données métabolomiques et NIRS sur un grand nombre de génotypes et dans des conditions environnementales variées. On peut anticiper que ce large jeu de données multi-omiques ouvrira des perspectives sur leur intégration statistique en vue d'améliorer la prédiction et la compréhension des mécanismes impliqués dans l'adaptation de la vigne aux contraintes exercées par le changement climatique. Pour traiter de ces questions, je pourrai poursuivre mes collaborations avec Leopoldo Sanchez et Harold Duruflé de mon ancienne unité et je pourrai par ailleurs profiter d'une certaine dynamique multi-espèces sur ce sujet au sein de l'UMR AGAP Institut en particulier mais aussi au niveau national via la métaprogramme INRAE DIGIT-BIO. Renaud Rincet, avec qui j'avais fait la preuve de concept de la prédiction phénotypique, est également intéressé par ce sujet d'intégration multi-omiques, et il pourra ainsi être un partenaire de choix pour mener à bien cette partie.

Pour conclure, je voudrai revenir sur les motivations qui m'ont poussées à présenter ce mémoire HDR. Il s'agit là notamment de concrétiser les encadrements passés et en cours, mais aussi de pouvoir permettre à mon équipe de conserver une certaine capacité d'accueil de doctorants, malgré les départs à la retraite (passés et à venir) de chercheurs confirmés. Au delà de ces aspects, et si cela a nécessité un certain investissement personnel, je me rends compte que cet exercice a été particulièrement intéressant, puisqu'il m'a permis de réaliser un bilan réflexif sur ma carrière et de poser mes perspectives de recherche à différentes échéances.

Bibliographie

- E. Albert, V. Segura, J. Gricourt, et al. Association mapping reveals the genetic architecture of tomato response to water deficit : Focus on major fruit quality traits. *Journal of Experimental Botany*, 67(22) :6413–6430, 2016. doi : 10.1093/jxb/erw411.
- M. R. Allwright, A. Payne, G. Emiliani, et al. Biomass traits and candidate genes for bioenergy revealed through association genetics in coppiced European *Populus nigra* (L.). *Biotechnology for Biofuels*, 9(1) :195, 2016. doi : 10.1186/s13068-016-0603-1.
- B. Arouisse, A. Korte, F. van Eeuwijk, and W. Kruijer. Imputation of 3 million SNPs in the Arabidopsis regional mapping population. *The Plant Journal*, 102(4) :872–882, 2020. doi : 10.1111/tpj.14659.
- W. Astle and D. J. Balding. Population Structure and Cryptic Relatedness in Genetic Association Studies. *Statistical Science*, 24(4) :451–471, 2009. doi : 10.1214/09-STS307.
- S. Atwell, Y. S. Huang, B. J. Vilhjálmsson, et al. Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. *Nature*, 465(7298) :627–631, 2010. doi : 10.1038/nature08800.
- D. J. Balding. A tutorial on statistical methods for population association studies. *Nature Reviews Genetics*, 7(10) :781–791, 2006. doi : 10.1038/nrg1916.
- G. Bauchet, S. Grenier, N. Samson, et al. Identification of major loci and genomic regions controlling acid and volatile content in tomato fruit : Implications for flavor improvement. *New Phytologist*, 215(2) :624–641, 2017. doi : 10.1111/nph.14615.
- K. A. Bindon, P. R. Dry, and B. R. Loveys. Influence of Plant Water Status on the Production of C₁₃-Norisoprenoid Precursors in *Vitis vinifera* L. Cv. Cabernet Sauvignon Grape Berries. *Journal of Agricultural and Food Chemistry*, 55(11) : 4493–4500, 2007. doi : 10.1021/jf063331p.
- F. Bonnafous, A. Duhnen, L. Gody, et al. Mlmm.gwas : Pipeline for GWAS Using MLMM, 2019.

- J.-M. Boursiquot, C. Marchal, and T. Lacombe. Panel-C4, a grapevine core collection designed for climate change studies. 12. International Conference on Grapevine Breeding and Genetics, 2018.
- E. A. Boyle, Y. I. Li, and J. K. Pritchard. An Expanded View of Complex Traits : From Polygenic to Omnigenic. *Cell*, 169(7) :1177–1186, 2017. doi : 10.1016/j.cell.2017.05.038.
- C. Brault, V. Segura, P. This, et al. Across-population genomic prediction in grapevine opens up promising prospects for breeding. *bioRxiv*, soumis. doi : 10.1101/2021.07.29.454290.
- D. G. Butler, B. R. Cullis, A. R. Gilmour, et al. ASReml-R Reference Manual, 2018.
- L. R. Cardon and L. J. Palmer. Population stratification and spurious allelic association. *The Lancet*, 361(9357) :598–604, 2003. doi : 10.1016/S0140-6736(03)12520-2.
- S. D. Castellarin, M. A. Matthews, G. Di Gaspero, and G. A. Gambetta. Water deficits accelerate ripening and induce changes in gene expression regulating flavonoid biosynthesis in grape berries. *Planta*, 227(1) :101–112, 2007a. doi : 10.1007/s00425-007-0598-8.
- S. D. Castellarin, A. Pfeiffer, P. Sivilotti, et al. Transcriptional regulation of anthocyanin biosynthesis in ripening fruits of grapevine under seasonal water deficit. *Plant, Cell & Environment*, 30(11) :1381–1399, 2007b. doi : 10.1111/j.1365-3040.2007.01716.x.
- A. Chateigner, M.-C. Lesage-Descauses, O. Rogier, et al. Gene expression predictions and networks in natural populations supports the omnigenic theory. *BMC Genomics*, 21(1) :416, 2020. doi : 10.1186/s12864-020-06809-2.
- M. M. Chaves, J. M. Costa, O. Zarrouk, et al. Controlling stomatal aperture in semi-arid regions—The dilemma of saving water or being cool? *Plant Science*, 251 :54–64, 2016. doi : 10.1016/j.plantsci.2016.06.015.
- J. Chen and Z. Chen. Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3) :759–771, 2008. doi : 10.1093/biomet/asn034.
- P. J. Conner, S. K. Brown, and N. F. Weeden. Molecular-marker analysis of quantitative traits for growth and development in juvenile apple trees. *Theoretical and Applied Genetics*, 96(8) :1027–1035, 1998. doi : 10.1007/s001220050835.
- A. Coupel-Ledru, E. Lebon, A. Christophe, et al. Reduced nighttime transpiration is a relevant breeding target for high water-use efficiency in grapevine. *Proceedings of the National Academy of Sciences*, 113(32) :8963–8968, 2016. doi : 10.1073/pnas.1600826113.

- A. Coupel-Ledru, B. Pallas, M. Delalande, et al. Tree architecture, light interception and water use related traits are controlled by different genomic regions in an apple tree core collection. *New Phytologist*, soumis.
- G. Covarrubias-Pazaran. Software update : Moving the R package sommer to multivariate mixed models for genome-assisted prediction. *bioRxiv*, – –, 2018.
- I. De Wit, N. Cook, and J. Keulemans. Characterization of tree architecture in two-year-old apple seedling populations of different progenies with a common columnar gene parent. In *Acta Horticulturae*, volume 663, pages 363–368, 2004. doi : 10.17660/ActaHortic.2004.663.62.
- L. G. Deluc, D. R. Quilici, A. Decendit, et al. Water deficit alters differentially metabolic pathways affecting important flavor and quality traits in grape berries of Cabernet Sauvignon and Chardonnay. *BMC Genomics*, 10(1) :212, 2009. doi : 10.1186/1471-2164-10-212.
- B. Devlin and K. Roeder. Genomic control for association studies. *Biometrics*, 55(4) :997–1004, 1999. doi : 10.1111/j.0006-341x.1999.00997.x.
- E. Duchêne. How can grapevine genetics contribute to the adaptation to climate change ? *OENO One*, 50(3), 2016. doi : 10.20870/oeno-one.2016.50.3.98.
- R. J. Elshire, J. C. Glaubitz, Q. Sun, et al. A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. *PLOS ONE*, 6(5) :e19379, 2011. doi : 10.1371/journal.pone.0019379.
- A. M. Fahrenkrog, L. G. Neves, M. F. R. Resende Jr, et al. Genome-wide association study reveals putative regulators of bioenergy traits in *Populus deltoides*. *New Phytologist*, 213(2) :799–811, 2017. doi : 10.1111/nph.14154.
- P. Faivre-Rampant, G. Zaina, V. Jorge, et al. New resources for genetic studies in *Populus nigra* : Genome-wide SNP discovery and development of a 12k Infinium array. *Molecular Ecology Resources*, 16(4) :1023–1036, 2016. doi : 10.1111/1755-0998.12513.
- R. A. Fisher. The Correlation between Relatives on the Supposition of Mendelian Inheritance. *Earth and Environmental Science Transactions of The Royal Society of Edinburgh*, 52(2) :399–433, 1919. doi : 10.1017/S0080456800012163.
- T. Flutre, L. Le Cunff, A. Fodor, et al. Genome-wide association and prediction studies using a grapevine diversity panel give insights into the genetic architecture of several traits of interest. *bioRxiv*, 2020. doi : 10.1101/2020.09.10.290890.
- A. Fodor, V. Segura, M. Denis, et al. Genome-wide prediction methods in highly diverse and heterozygous species : Proof-of-concept through simulation in grapevine. *PloS One*, 9(11) :e110436, 2014. doi : 10.1371/journal.pone.0110436.

- R. J. Galán, A.-M. Bernal-Vasquez, C. Jebsen, et al. Hyperspectral Reflectance Data and Agronomic Traits Can Predict Biomass Yield in Winter Rye Hybrids. *BioEnergy Research*, 13(1) :168–182, 2020. doi : 10.1007/s12155-019-10080-z.
- G. A. Gambetta, J. C. Herrera, S. Dayer, et al. The physiology of drought stress in grapevine : Towards an integrative definition of drought tolerance. *Journal of Experimental Botany*, 71(16) :4658–4676, 2020. doi : 10.1093/jxb/eraa245.
- J. A. Gamon, J. Peñuelas, and C. B. Field. A narrow-waveband spectral index that tracks diurnal changes in photosynthetic efficiency. *Remote Sensing of Environment*, 41(1) :35–44, 1992. doi : 10.1016/0034-4257(92)90059-S.
- M. N. Gebreselassie, K. Ader, N. Boizot, et al. Near-infrared spectroscopy enables the genetic analysis of chemical properties in a large set of wood samples from *Populus nigra* (L.) natural populations. *Industrial Crops and Products*, 107 : 159–171, 2017. doi : 10.1016/j.indcrop.2017.05.013.
- GIEC. Résumé à l'intention des décideurs, Changements climatiques 2013 : Les éléments scientifiques, 2013.
- H. Giraud, C. Lehermeier, E. Bauer, et al. Linkage Disequilibrium with Linkage Analysis of Multiline Crosses Reveals Different Multi-allelic QTL for Hybrid Performance in the Flint and Dent Heterotic Groups of Maize. *Genetics*, 198(4) : 1717–1734, 2014. doi : 10.1534/genetics.114.169367.
- I. A. Graham, K. Besser, S. Blumer, et al. The Genetic Map of *Artemisia annua* L. Identifies Loci Affecting Yield of the Antimalarial Drug Artemisinin. *Science*, 327(5963) :328–331, 2010. doi : 10.1126/science.1182612.
- M. Griesser, G. Weingart, K. Schoedl-Hummel, et al. Severe drought stress is affecting selected primary metabolites, polyphenols, and volatile metabolites in grapevine leaves (*Vitis vinifera* cv. Pinot noir). *Plant Physiology and Biochemistry*, 88 :17–26, 2015. doi : 10.1016/j.plaphy.2015.01.004.
- M. Grzybowski, N. K. Wijewardane, A. Atefi, et al. Hyperspectral reflectance-based phenotyping for quantitative genetics in crops : Progress and challenges. *Plant Communications*, 2(4) :100209, 2021. doi : 10.1016/j.xplc.2021.100209.
- F. P. Guerra, J. L. Wegrzyn, R. Sykes, et al. Association genetics of chemical wood properties in black poplar (*Populus nigra*). *New Phytologist*, 197(1) :162–176, 2013. doi : 10.1111/nph.12003.
- C. R. Henderson. Applications of linear models in animal breeding. *undefined*, 1984.
- N. Heslot, D. Akdemir, M. E. Sorrells, and J.-L. Jannink. Integrating environmental covariates and crop modeling into the genomic selection framework to predict genotype by environment interactions. *Theoretical and Applied Genetics*, 127(2) : 463–480, 2014. doi : 10.1007/s00122-013-2231-5.

- U. Hochberg, A. Degu, G. R. Cramer, et al. Cultivar specific metabolic changes in grapevines berry skins in relation to deficit irrigation and hydraulic behavior. *Plant Physiology and Biochemistry*, 88 :42–52, 2015. doi : 10.1016/j.plaphy.2015.01.006.
- M. W. Horton, A. M. Hancock, Y. S. Huang, et al. Genome-wide patterns of genetic variation in worldwide *Arabidopsis thaliana* accessions from the RegMap panel. *Nature Genetics*, 44(2) :212–216, 2012. doi : 10.1038/ng.1042.
- M. Huang, X. Liu, Y. Zhou, et al. BLINK : A package for the next level of genome-wide association studies with both individuals and markers in the millions. *GigaScience*, 8(2), 2019. doi : 10.1093/gigascience/giy154.
- Y. Itoh and Y. Yamada. Relationships between genotype x environment interaction and genetic correlation of the same trait measured in different environments. *Theoretical and Applied Genetics*, 80(1) :11–16, 1990. doi : 10.1007/BF00224009.
- R. C. Jansen. Interval mapping of multiple quantitative trait loci. *Genetics*, 135(1) :205–211, 1993. doi : 10.1093/genetics/135.1.205.
- D. Jarquín, J. Crossa, X. Lacaze, et al. A reaction norm model for genomic selection using high-dimensional genomic and environmental data. *Theoretical and Applied Genetics*, 127(3) :595–607, 2014. doi : 10.1007/s00122-013-2243-1.
- J. M. Jez and J. P. Noel. Reaction Mechanism of Chalcone Isomerase. *Journal of Biological Chemistry*, 277(2) :1361–1369, 2002. doi : 10.1074/jbc.M109224200.
- H. M. Kang, J. H. Sul, S. K. Service, et al. Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics*, 42(4) :348–354, 2010. doi : 10.1038/ng.548.
- A. Korte, B. J. Vilhjálmsson, V. Segura, et al. A mixed-model approach for genome-wide association studies of correlated traits in structured populations. *Nature Genetics*, 44(9) :1066–1071, 2012. doi : 10.1038/ng.2376.
- M. R. Krause, L. González-Pérez, J. Crossa, et al. Hyperspectral Reflectance-Derived Relationship Matrices for Genomic Prediction of Grain Yield in Wheat. *G3 Genes/Genomes/Genetics*, 9(4) :1231–1247, 2019. doi : 10.1534/g3.118.200856.
- C. Lafon-Placette, A.-L. Le Gac, D. Chauveau, et al. Changes in the epigenome and transcriptome of the poplar shoot apical meristem in response to water availability affect preferentially hormone pathways. *Journal of Experimental Botany*, 69(3) :537–551, 2018. doi : 10.1093/jxb/erx409.
- H. M. Lane, S. C. Murray, O. A. Montesinos-López, et al. Phenomic selection and prediction of maize grain yield from near-infrared reflectance spectroscopy of kernels. *The Plant Phenome Journal*, 3(1) :e20002, 2020. doi : 10.1002/ppj2.20002.

- H. Lango Allen, K. Estrada, G. Lettre, et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*, 467(7317) : 832–838, 2010. doi : 10.1038/nature09410.
- A.-L. Le Gac, C. Lafon-Placette, D. Chauveau, et al. Winter-dormant shoot apical meristem in poplar trees shows environmental epigenetic memory. *Journal of Experimental Botany*, 69(20) :4821–4837, 2018. doi : 10.1093/jxb/ery271.
- X. Liu, M. Huang, B. Fan, et al. Iterative Usage of Fixed and Random Effect Models for Powerful and Efficient Genome-Wide Association Studies. *PLOS Genetics*, 12(2) :e1005767, 2016. doi : 10.1371/journal.pgen.1005767.
- L. Ma, S. Han, J. Yang, and Y. Da. Multi-locus Test Conditional on Confirmed Effects Leads to Increased Power in Genome-wide Association Studies. *PLOS ONE*, 5(11) :e15006, 2010. doi : 10.1371/journal.pone.0015006.
- T. A. Manolio, F. S. Collins, N. J. Cox, et al. Finding the missing heritability of complex diseases. *Nature*, 461(7265) :747–753, 2009. doi : 10.1038/nature08494.
- C. Marfil, V. Ibañez, R. Alonso, et al. Changes in grapevine DNA methylation and polyphenols content induced by solar ultraviolet-B radiation, water deficit and abscisic acid spray treatments. *Plant physiology and biochemistry : PPB*, 135 : 287–294, 2019. doi : 10.1016/j.plaphy.2018.12.021.
- Z. Migicovsky, J. Sawler, K. M. Gardner, et al. Patterns of genomic and phenomic diversity in wine and table grapes. *Horticulture Research*, 4(1) :1–11, 2017. doi : 10.1038/hortres.2017.35.
- F. Muñoz and L. Sanchez. *breedR : Statistical Methods for Forest Genetic Resources Analysts*, 2021.
- S. D. Nicolas, J.-P. Péros, T. Lacombe, et al. Genetic diversity, linkage disequilibrium and power of a large grapevine (*Vitis vinifera* L) diversity panel newly designed for association studies. *BMC Plant Biology*, 16(1) :74, 2016. doi : 10.1186/s12870-016-0754-z.
- N. Ollat and J.-M. Touzard. La vigne, le vin, et le changement climatique en France - Projet LACCAVE - Horizon 2050, 2020.
- M. Pégard, V. Segura, F. Muñoz, et al. Favorable Conditions for Genomic Evaluation to Outperform Classical Pedigree Evaluation Highlighted by a Proof-of-Concept Study in Poplar. *Frontiers in Plant Science*, 11 :581954, 2020. doi : 10.3389/fpls.2020.581954.
- A. Persoons, A. Maupetit, C. Louet, et al. Genomic signatures of a major adaptive event in the pathogenic fungus *Melampsora larici-populina*. *bioRxiv*, soumis. doi : 10.1101/2021.04.09.439223.

- L. Pinasseau, A. Vallverdú-Queralt, A. Verbaere, et al. Cultivar Diversity of Grape Skin Polyphenol Composition and Changes in Response to Drought Investigated by LC-MS Based Metabolomics. *Frontiers in Plant Science*, 8 :1826, 2017. doi : 10.3389/fpls.2017.01826.
- A. Platt, B. J. Vilhjálmsson, and M. Nordborg. Conditions Under Which Genome-Wide Association Studies Will be Positively Misleading. *Genetics*, 186(3) :1045–1052, 2010. doi : 10.1534/genetics.110.121665.
- I. Porth, J. Klapšte, O. Skyba, et al. Genome-wide association mapping for wood characteristics in *Populus* identifies an array of candidate single nucleotide polymorphisms. *New Phytologist*, 200(3) :710–726, 2013. doi : 10.1111/nph.12422.
- A. L. Price, N. J. Patterson, R. M. Plenge, et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38 (8) :904–909, 2006. doi : 10.1038/ng1847.
- A. L. Price, N. A. Zaitlen, D. Reich, and N. Patterson. New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics*, 11(7) : 459–463, 2010. doi : 10.1038/nrg2813.
- J. K. Pritchard, M. Stephens, N. A. Rosenberg, and P. Donnelly. Association Mapping in Structured Populations. *The American Journal of Human Genetics*, 67 (1) :170–181, 2000. doi : 10.1086/302959.
- S. Pulkka, V. Segura, A. Harju, et al. Prediction of Stilbene Content from Heartwood Increment Cores of Scots Pine Using near Infrared Spectroscopy Methodology. *Journal of Near Infrared Spectroscopy*, 24(6) :517–528, 2016.
- A. M. Rae, N. R. Street, K. M. Robinson, et al. Five QTL hotspots for yield in short rotation coppice bioenergy poplar : The Poplar Biomass Loci. *BMC Plant Biology*, 9(1) :23, 2009. doi : 10.1186/1471-2229-9-23.
- R. Rincent, J.-P. Charpentier, P. Faivre-Rampant, et al. Phenomic Selection Is a Low-Cost and High-Throughput Method Based on Indirect Predictions : Proof of Concept on Wheat and Poplar. *G3 Genes/Genomes/Genetics*, 8(12) :3961–3972, 2018. doi : 10.1534/g3.118.200760.
- O. Rogier, A. Chateigner, S. Amanzougarene, et al. Accuracy of RNAseq based SNP discovery and genotyping in *Populus nigra*. *BMC Genomics*, 19(1) :909, 2018. doi : 10.1186/s12864-018-5239-z.
- S. Y. Rogiers, Z. A. Coetzee, R. R. Walker, et al. Potassium in the Grape (*Vitis vinifera* L.) Berry : Transport and Function. *Frontiers in Plant Science*, 8 :1629, 2017. doi : 10.3389/fpls.2017.01629.
- C. Sauvage, V. Segura, G. Bauchet, et al. Genome-Wide Association in Tomato Reveals 44 Candidate Loci for Fruit Metabolic Traits. *Plant Physiology*, 165(3) : 1120–1132, 2014. doi : 10.1104/pp.114.241521.

- S. Savoi, D. C. J. Wong, P. Arapitsas, et al. Transcriptome and metabolite profiling reveals that prolonged drought modulates the phenylpropanoid and terpenoid pathway in white grapes (*Vitis vinifera* L.). *BMC Plant Biology*, 16(1) :67, 2016. doi : 10.1186/s12870-016-0760-1.
- V. Segura, C. Cilas, F. Laurens, and E. Costes. Phenotyping progenies for complex architectural traits : A strategy for 1-year-old apple trees (*Malus x domestica* Borkh.). *Tree Genetics & Genomes*, 2(3) :140–151, 2006. doi : 10.1007/s11295-006-0037-1.
- V. Segura, C. Denancé, C.-E. Durel, and E. Costes. Wide range QTL analysis for complex architectural traits in a 1-year-old apple progeny. *Genome*, 50(2) : 159–171, 2007. doi : 10.1139/G07-002.
- V. Segura, C. Cilas, and E. Costes. Dissecting apple tree architecture into genetic, ontogenetic and environmental effects : Mixed linear modelling of repeated spatial and temporal measures. *New Phytologist*, 178(2) :302–314, 2008a. doi : 10.1111/j.1469-8137.2007.02374.x.
- V. Segura, A. Ouangraoua, P. Ferraro, and E. Costes. Comparison of tree architecture using tree edit distances : Application to 2-year-old apple hybrids. *Euphytica*, 161(1) :155–164, 2008b. doi : 10.1007/s10681-007-9430-6.
- V. Segura, C.-E. Durel, and E. Costes. Dissecting apple tree architecture into genetic, ontogenetic and environmental effects : QTL mapping. *Tree Genetics & Genomes*, 5(1) :165–179, 2009. doi : 10.1007/s11295-008-0181-x.
- V. Segura, B. J. Vilhjálmsson, A. Platt, et al. An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nature Genetics*, 44(7) :825–830, 2012. doi : 10.1038/ng.2314.
- A. S. Sergent, V. Segura, J. P. Charpentier, et al. Assessment of resistance to xylem cavitation in cordilleran cypress using near-infrared spectroscopy. *Forest Ecology and Management*, 462 :117943, 2020. doi : 10.1016/j.foreco.2020.117943.
- T. Simonneau, N. Ollat, A. Pellegrino, and E. Lebon. Contrôle de l'état hydrique dans la plante et réponses physiologiques de la vigne à la contrainte hydrique. *Innovations Agronomiques*, 38 :13–32, 2014.
- M. D. Sow, V. Segura, S. Chamaillard, et al. Narrow-sense heritability and PST estimates of DNA methylation in three *Populus nigra* L. populations under contrasting water availability. *Tree Genetics & Genomes*, 14(5) :78, 2018. doi : 10.1007/s11295-018-1293-6.
- M. D. Sow, A.-L. Le Gac, R. Fichot, et al. RNAi suppression of DNA methylation affects the drought stress response and genome integrity in transgenic poplar. *New Phytologist*, 232(1) :80–97, 2021. doi : 10.1111/nph.17555.

- J. Tello, C. Roux, H. Chouiki, et al. A novel high-density grapevine (*Vitis vinifera* L.) integrated linkage map using GBS in a half-diallel population. *Theoretical and Applied Genetics*, 132(8) :2237–2252, 2019. doi : 10.1007/s00122-019-03351-y.
- C. J. Tucker. Red and photographic infrared linear combinations for monitoring vegetation. *Remote Sensing of Environment*, 8(2) :127–150, 1979. doi : 10.1016/0034-4257(79)90013-0.
- A. R. Wade, H. Duruflé, L. Sanchez, and V. Segura. eQTLs are key players in the integration of genomic and transcriptomic data for phenotype prediction. *bioRxiv*, soumis. doi : 10.1101/2021.09.07.459279.
- J. L. Wegrzyn, A. J. Eckert, M. Choi, et al. Association genetics of traits controlling lignin and cellulose biosynthesis in black cottonwood (*Populus trichocarpa*, Salicaceae) secondary xylem. *New Phytologist*, 188(2) :515–532, 2010. doi : 10.1111/j.1469-8137.2010.03415.x.
- J. Yang, N. A. Zaitlen, M. E. Goddard, et al. Advantages and pitfalls in the application of mixed-model association methods. *Nature Genetics*, 46(2) :100–106, 2014. doi : 10.1038/ng.2876.
- J. Yu, G. Pressoir, W. H. Briggs, et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics*, 38(2) :203–208, 2006. doi : 10.1038/ng1702.
- Z. B. Zeng. Precision mapping of quantitative trait loci. *Genetics*, 136(4) :1457–1468, 1994. doi : 10.1093/genetics/136.4.1457.
- K. Zhao, M. J. Aranzana, S. Kim, et al. An Arabidopsis Example of Association Mapping in Structured Samples. *PLOS Genetics*, 3(1) :e4, 2007. doi : 10.1371/journal.pgen.0030004.
- X. Zhou and M. Stephens. Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nature Methods*, 11(4) :407–409, 2014. doi : 10.1038/nmeth.2848.
- X. Zhu, H. P. Maurer, M. Jenz, et al. The performance of phenomic selection depends on the genetic architecture of the target trait. *Theoretical and Applied Genetics*, 2021. doi : 10.1007/s00122-021-03997-7.

Annexes :

**Tirés à part des principaux travaux
scientifiques, classés par ordre chronologique.**



Dissecting apple tree architecture into genetic, ontogenetic and environmental effects: mixed linear modelling of repeated spatial and temporal measures

Vincent Segura¹, Christian Cilas² and Evelyne Costes¹

¹INRA, UMR DAP, INRA – Montpellier SupAgro – CIRAD – Université Montpellier II, Equipe Architecture et Fonctionnement des Espèces Fruitières, 2 place P. Viala, 34060 Montpellier Cedex 1, France; ²CIRAD, CP, TA 80/02, Avenue Agropolis, 34398 Montpellier Cedex 5, France

Summary

Author for correspondence:

E. Costes

Tel: +33 4 99612515

Fax: +33 4 99612616

Email: costes@supagro.inra.fr

Received: 1 November 2007

Accepted: 17 December 2007

- The present study aimed to dissect tree architectural plasticity into genetic, ontogenetic and environmental effects over the first 4 yr of growth of an apple (*Malus × domestica*) F1 progeny by means of mixed linear modelling of repeated data.
- Traits related to both growth and branching processes were annually assessed on different axes of the trees planted in a staggered-start design. Both spatial repetitions, (i.e. different axis types) and temporal repetitions (i.e. successive ages of trees) were considered in a mixed linear model of repeated data.
- A significant genotype effect was found for most studied traits and interactions between genotype and year and/or age were also detected. The analysis of repeated temporal measures highlighted that the magnitude of the decrease in primary growth is mainly determined by the first year of growth, and the decrease in bottom diameter increment is concomitant with the first fruiting occurrence.
- This approach allowed us to distinguish among the traits that were under genetic control, those for which this control is exerted differentially throughout tree life or depending on climatic conditions or an axis type. Mapping quantitative trait loci (QTL) that are specific to these different effects will constitute the next step in the research.

Key words: autocorrelation, branching, *Malus × domestica* (apple), phenotypic plasticity, primary growth, secondary growth.

New Phytologist (2008) **178**: 302–314

© The Authors (2008). Journal compilation © *New Phytologist* (2008)

doi: 10.1111/j.1469-8137.2008.02374.x

Introduction

Controlling plant architecture is often a desirable goal for perennial crop species and many studies have been performed to analyse the genetic determinism of tree growth and branching (Bradshaw & Stettler, 1995; Plomion *et al.*, 1996; Wu & Stettler, 1996, 1998; Wu, 1998; Scotti-Saintagne *et al.*, 2004). In fruit trees, architectural traits are generally controlled by using size-controlling rootstock, pruning and training (Costes *et al.*, 2006). However, a deeper understanding of the genetic determinism of fruit tree architecture could allow growth and branching traits to be introduced into selection schemes and thus reduce the pruning and training costs (Laurens *et al.*, 2000).

A substantial architectural variability has long been described in apple species and had led to the classification of cultivars into architectural types (Lespinasse, 1977, 1992). Over the last decade, several genetic studies have been performed on architectural traits in apple tree. One of our studies showed that many traits had fairly high broad-sense heritability in a 1-yr-old apple progeny, especially branching traits or primary growth traits such as mean internode length (Segura *et al.*, 2006, 2007). Quantitative trait loci (QTLs) have also been mapped for these traits in juvenile progenies (Conner *et al.*, 1998; Liebhard *et al.*, 2003; Kenis & Keulemans, 2007; Segura *et al.*, 2007), but when the results of these studies are compared, heritability estimates and QTLs mapped appear to be unstable. This was first explained by the variability of the

genetic background analysed in these studies, but when the same progeny was studied over several years, variability was again seen in heritability estimates and QTLs mapped (Liebhard *et al.*, 2003; Kenis & Keulemans, 2007). Similar results were also found for growth traits assessed in forest trees such as poplar (Bradshaw & Stettler, 1995; Wu & Stettler, 1996; Wu, 1998). Moreover, when different axes in a tree were studied at the same age, variability in genetic determinism was again observed between axes; and, for example, in a previous study we detected distinct QTLs between the trunk and branches for internode lengthening in apple hybrids (Segura *et al.*, 2007). These results underline the marked plasticity of architectural traits in trees. Indeed, like most traits that are important for fitness and agricultural value, architectural traits are complex and greatly influenced by environmental factors, such as soil characteristics and temperature or water resources. This means that the observed phenotype results from genes, environmental factors and their interactions (Lynch & Walsh, 1998; Holland, 2007). It therefore follows that genotype \times environment interactions have been of great interest for years in plant science (Comstock & Moll, 1963), and many studies have been performed in this field, especially on annual crops. However, when considering tree species, few studies have dealt with genotype \times environment interactions (Osorio *et al.*, 2001; Finn *et al.*, 2003; Legave *et al.*, 2006; Sykes *et al.*, 2006). To assess these interactions, specific designs need to be devised, for example multiple location designs, which are usually very expensive and time-consuming, especially for perennial species.

In addition to environmental factors which confer a phenotypic plasticity by their interactions with genes, other factors affect the growth of a tree throughout its life. Even though tree structure often results from repetitive processes (White, 1979), the successive growth units are not totally similar as they are related to morphogenetic gradients during tree ontogeny (Barthélémy *et al.*, 1997). This concept of morphogenetic gradient states that bud fate changes according to position within the tree structure and during plant development, and thus integrates both temporal and spatial architectural plasticity observed within a tree structure. One of the most evident morphogenetic gradients is the decrease in height increment with tree age, also called age-related decline in growth. In forest trees, many studies have investigated this decline in primary growth (see Bond, 2000 for a review; Bond *et al.*, 2007), and genetic characterization of ontogenetic effects was recently achieved through functional mapping in poplar (Wu *et al.*, 2003, 2004; Wu & Lin, 2006; Yang *et al.*, 2006), highlighting the differential expression of genes during tree life. In apple species, the decline in primary growth with tree age has been demonstrated in different cultivars (Costes *et al.*, 2003) and for different rootstocks (Seleznyova *et al.*, 2003). In addition, stochastic approaches have highlighted ontogenetic gradients for branching patterns on the trunk and growth units of apple cultivars (Costes & Guédon, 2002;

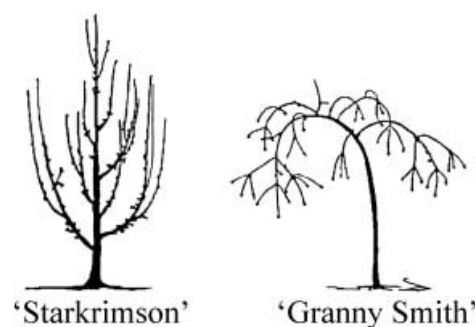


Fig. 1 Schematic representation of the architecture of the two parent genotypes 'Starkrimson' and 'Granny Smith'. (Source: Lespinasse (1992).)

Renton *et al.*, 2006). However, even though these ontogenetic gradients have been shown in apple cultivars, they have not yet been clearly characterized particularly because ontogenetic and environmental (e.g. at least climatic) effects were often merged in genetic studies.

Growth and branching traits are affected by genes, environment and ontogeny, and their interaction, and our knowledge of these factors is still limited. The following questions are thus addressed in apple species: what is the relative contribution made by genes, environment and their interaction to architectural traits; how can ontogenetic gradients that affect tree architecture be characterized from a genetic point of view? On the basis of a specific staggered-start mating design (Loughin, 2006), the present study aimed to investigate genetic, environmental and ontogenetic effects on architectural traits in apple species. To do this, mixed linear models of repeated data were used for primary growth, secondary growth and branching traits assessed over the first 4 yr of growth in apple tree hybrids.

Materials and Methods

Plant material

The studied apple (*Malus \times domestica* Borkh.) F1 progeny was derived from a 'Starkrimson' \times 'Granny Smith' cross, with parents being chosen for their contrasting architecture (Fig. 1). The 'Starkrimson' maternal parent displayed an erect growth habit with many short shoots and a tendency to irregular bearing (Lespinasse, 1992). By contrast, the 'Granny Smith' pollen parent displayed a weeping habit with long shoots and fruit-bearing regularity.

This F1 progeny comprised 125 seedlings that were grown on their own roots in a nursery in 2002. One year later, graft wood was taken when possible from three successive nodes in the middle of main axes for 50 genotypes randomly selected, to produce replicates. The 150 trees obtained were planted in March 2003 at the Melgueil INRA Montpellier experimental

Axes	Design	Years of growth	Tree age	Annual shoots assessed
Trunk and LSAS	A	2003 to 2006	1 to 4	AS1 to AS4
	B	2004 to 2006	1 to 3	AS1 to AS3
LPAS	A	2004 to 2006	2 to 4	AS1 to AS3
	B	2005 to 2006	2 to 3	AS1 to AS2

Table 1 Annual shoots (AS) assessed on trunk, long sylleptic axillary shoot (LSAS) and long proleptic axillary shoot (LPAS) of trees of designs A and B (tree age and AS years of growth are also indicated)

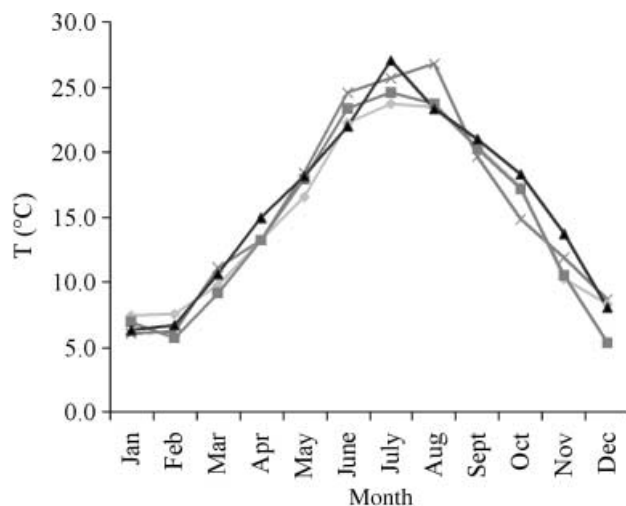


Fig. 2 Average monthly temperature in 2003 (crosses), 2004 (diamonds), 2005 (squares) and 2006 (triangles). Data are from the meteorological station of Météo-France, Fréjorgues, Montpellier, France.

station (France) 5 × 2 m apart in an east–west orientation in 10-tree microplots randomly scattered throughout the field. At the same time, seedlings from the nursery were cut back and transplanted to the same site. One year later, graft wood was taken on seedling trees from two successive nodes in the middle of long sylleptic axillary shoots (LSAS) on the 123 seedling trees (two trees died). These shoots were selected among the longest ones within the trees (born in the middle part of the trunks) and were composed of > 20 internodes. Two grafts were carried out for each of the 123 genotypes to produce replicates. The 246 trees obtained were planted in March 2004 at the Melgueil INRA Montpellier experimental station 5 × 1.5 m apart in an east–west orientation in six-tree microplots randomly scattered throughout the field. In both cases, graftings were performed onto ‘Pajam 1’ rootstock, a clonal selection of M9, which confers low vigour, a short juvenile period and substantial, regular productivity. In order to study their architecture, all the trees were grown with minimal training, that is, they were not pruned and the trunks were staked up to 1 m. They were regularly irrigated using a microjet system to avoid soil water deficits. Pests and diseases were controlled by conventional means in line with professional

practices throughout this study. To account for the influence of fruiting occurrence on tree ontogeny, fruits were not thinned throughout the study and they were annually harvested. In such a specific mating design, also called a staggered-start design, some genotypes (50) were planted at the same site with an interval of 1 yr, and as a result for these 50 genotypes the same years of growth occurred during different climatic years. As trees were regularly irrigated, major differences between climatic years were in relation to temperature and these are illustrated in Fig. 2.

Architectural description

Trees were observed from 2004 to 2007, that is, the first 4 and 3 yr of growth were assessed in trees planted in 2003 and 2004, respectively. In both cases, observations were performed on the following sampled axes: the trunk, two LSAS and two long proleptic axillary shoots (LPAS), when present. LSAS grew on the first annual shoots (AS) of the trunk from the first year of growth, while LPAS also grew on the first AS of the trunk, but 1 yr later (from the second year of growth). These shoots were selected among the longest developed within each tree. They were always longer than 20 cm and composed at least of > 10 internodes. Measurements were performed on AS throughout different calendar years and tree ages for each design and axis type (e.g. trunk, LSAS, LPAS; see details in Table 1). They focused on three main processes: primary growth, including length, internode number and length of the longest internode (Fig. 3a); secondary growth, including bottom and top diameters (Fig. 3b); and branching, that is, number of axillary shoots, including both sylleptic and proleptic shoots (Fig. 3a). To provide descriptors as close as possible to biological processes, other traits were computed from measures such as mean internode length, bottom diameter increment, number of latent buds and percentage of branching nodes. Abbreviations and trait formula used in this study are indicated in Table 2.

Statistical analysis

Phenotyping data were gathered by trait on each axis type. First, only the 50 genotypes in the design planted in 2004, and corresponding to the 50 genotypes in the design planted in 2003, were considered. Second, all the data for the design

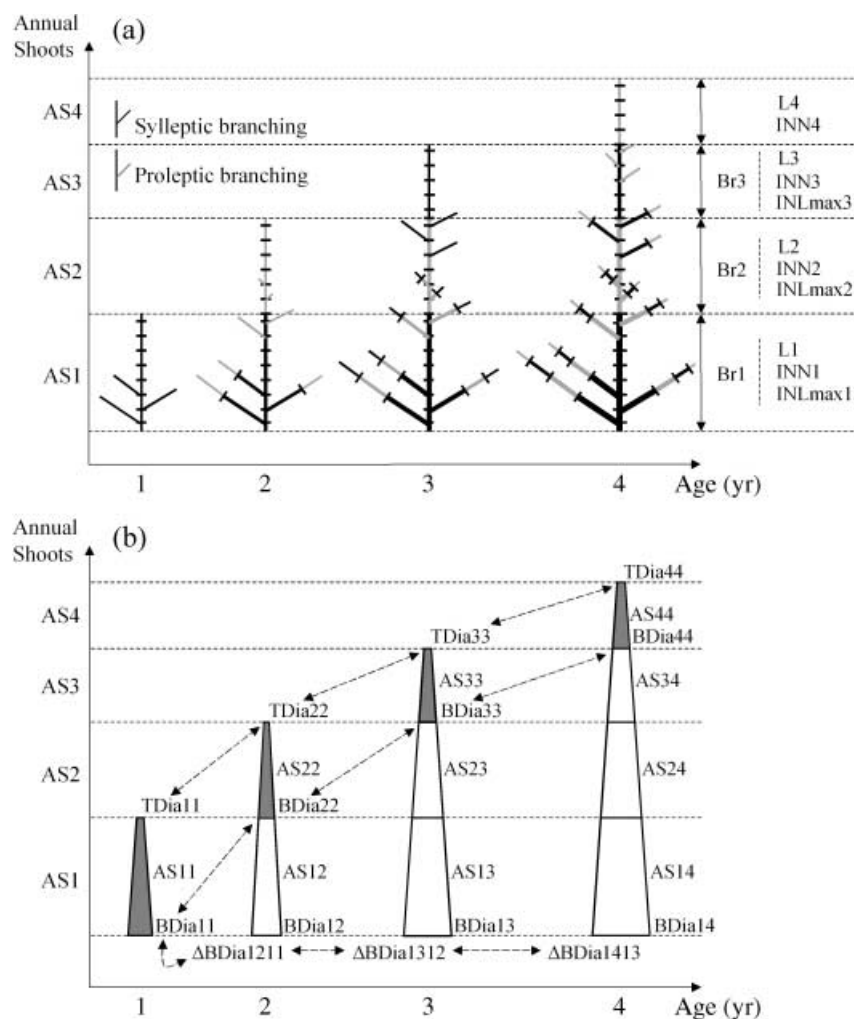


Fig. 3 Schematic representation of changes in architectural traits along an axis (trunk, long sylleptic axillary shoots (LSAS) or proleptic axillary shoots (LPAS)) phenotyped over the first 4 yr of growth. (a) Primary growth and branching; (b) secondary growth. For trait abbreviations, see Table 2.

Table 2 List of traits classified by primary growth, secondary growth or branching

Trait	Abbreviation	Formula
<i>Primary growth</i>		
Length	L	
Internode number	INN	
Mean internode length	INL	L/INN
Length of the longest internode	INLmax	
<i>Secondary growth</i>		
Bottom diameter increment	$\Delta Bdia$	$Bdia_{j+1} - Bdia_j^a$
Bottom diameter	Bdia	
Top diameter	Tdia	
<i>Branching</i>		
Number of axillary shoots	Br	
Number of latent buds	Latents	$INN - Br$
Percentage of branching nodes	%Br	Br/INN

^a j and $j + 1$ refer to successive tree ages.

planted in 2004 (125 genotypes) were used for the analysis. Datasets with 50 and 125 genotypes are referred to as dataset 1 and dataset 2, and it is noteworthy that dataset 2 was unbalanced. Both datasets were analysed by mixed linear modelling of repeated data.

The term 'repeated data' usually refers to multiple measurements taken in a sequence on the same experimental unit, and repeated data analysis assumes a covariance between repeated measurements, that is, a covariance matrix structure is modelled for residuals (Littell *et al.*, 1998, 2000; Moser, 2004). The experimental unit considered here was the tree, and repeated measures were of two types, depending on the axis analysed. In all cases, tree age was considered as repeated data (temporal repetition). For the trunk, as only one axis was observed within a tree, temporal repetition was the only repeated data considered. For the LSAS and LPAS, as two axes were observed within a tree, a spatial repetition was also taken into account in the model. In addition, to evaluate within-tree variability, data on all axes were gathered (including trunk, LSAS and LPAS data) and thus constituted the whole-tree

scale. In this case, two repeated data were also considered, temporal repetition and spatial (within tree) repetition between the different axes.

Considering the trunk, the following mixed linear model was built for each studied trait:

$$y_{ijkl} = \mu + \gamma_i + \alpha_j + \tau_k + (\gamma\alpha)_{ij} + (\gamma\tau)_{ik} + \varepsilon_{ijkl}, \quad \text{Eqn 1}$$

where y_{ijkl} is the phenotypic value measured on tree l of genotype i at age j in year k ; μ is the overall mean; γ_i is the random effect of the genotype i ; α_j is the fixed effect of the age j ; τ_k is fixed effect in year k ; $(\gamma\alpha)_{ij}$ is the random interaction between genotype i and age j ; $(\gamma\tau)_{ik}$ is the random interaction between genotype i and year k ; and ε_{ijkl} is the random residual error for tree l of genotype i at age j in year k . Several covariance structures were tested to model residuals on each trait. First, an unstructured covariance matrix (UN) was considered in the model to compute variances and covariances among ages. Second, if the variances and covariances showed a possible structure, a structured covariance matrix was chosen and fitted to the data. The following covariance structures were tested depending on the trait analysed: variance component (VC, without repeated data, no covariance and homogenous variance between ages); banded main diagonal (UN(1), no covariance and heterogeneous variances between ages); first-order antedependence (ANTE(1), heterogeneous covariances and variances between ages); heterogeneous first-order autoregressive (ARH(1), heterogeneous variances and covariances decreasing with increasing lag between ages). More details on covariance structures can be found in the mixed procedure section of SAS[®] v.8 software documentation (SAS Institute Inc., 2000). Third, the covariance structure that minimized the BIC (Bayes Schwarz information criterion) was selected (Littell *et al.*, 1998, 2000). Fourth, the final model was selected by removing nonsignificant effects until the BIC was minimal with the covariance structure selected in the previous step.

Considering the other axes (LSAS and LPAS) and the whole-tree scale, a random effect ζ_{il} of tree replication l nested into genotype i was added into the previous mixed linear model built at the trunk scale:

$$y_{ijklm} = \mu + \gamma_i + \alpha_j + \tau_k + (\gamma\alpha)_{ij} + (\gamma\tau)_{ik} + \zeta_{il} + \varepsilon_{ijklm}. \quad \text{Eqn 2}$$

In this case, as two repeated measures were considered (temporal and spatial repetitions), the covariance structure modelled on residuals is a direct (Kronecker) product between two covariance matrices. As was done previously, several covariance structures were tested to model residuals on each trait: UN \otimes UN, direct product between two unstructured covariance matrices; UN \otimes CS, direct product between an unstructured covariance matrix and a compound symmetry covariance matrix (homogenous variance and covariance between ages); UN \otimes AR(1), direct product between an unstructured covariance matrix and

a first-order autoregressive structure covariance matrix (homogenous variance and covariances decreasing with increasing lag between ages). More details on covariance structures can be found in the mixed procedure section of SAS[®] v.8 software documentation (SAS Institute Inc., 2000). In the covariance matrix direct product, the first term was always a UN covariance matrix, while the second term was either a UN, CS or AR(1) covariance matrix. As a result, depending on the case, the temporal or spatial repetitions were either modelled by the first or the second term in the covariance matrix direct product to select the model that best fitted the data. The methodology already described was used for model selection.

In all cases, whatever the axis type or the scale, spatial and temporal correlations were computed respectively among axes or ages from variances and covariances as:

$$\rho_{np} = \sigma_{np} / \sqrt{\sigma_n^2 \sigma_p^2}, \quad \text{Eqn 3}$$

where ρ_{mn} is the correlation between spatial or temporal repetitions m and n ; σ_{mn} is the covariance between spatial or temporal repetitions m and n ; σ_m^2 is the variance at age m or for axis m ; and σ_n^2 is the variance at age n or for axis n . The significance of the correlations was deduced from the significance of the Wald Z -test computed on the corresponding covariances.

Mixed modelling was performed using the restricted maximum likelihood (REML) method in the mixed procedure of SAS[®] v.8 software (SAS Institute Inc., 2000). The REML method was chosen because it is considered the most suitable for unbalanced datasets (Dieters *et al.*, 1995).

Results

For all scales (whole tree, or axes), axis types (trunk, LSAS or LPAS) and datasets (1 or 2), most fitted models included genotype (G), tree age (A) and year (Y) effects that were significant or highly significant (Tables 3, 5 and 6). A and Y effects are illustrated through changes in the mean values of each trait over tree development and differences between the two designs, respectively (Supplementary material, Fig. S1). In most cases, UN covariance matrices provided the best fit either for temporal or spatial repetition based on BIC minimization. Furthermore, when datasets 1 and 2 were compared, the same models were selected and an increase in the genotypic effect was generally observed, probably because of the larger number of genotypes considered. The following subsections, organized by architectural process, provide results for additional effects included in the models selected; covariance structures; and correlations between repeated effects.

Primary growth

All the models selected for primary growth traits followed the general case with significant or highly significant G, A and Y effects (Table 3, Fig. S1a). For AS length and number of

Table 3 Significance of effects (**, highly significant, $P \leq 0.01$; *, significant, $P \leq 0.05$) in selected models according to Bayes Schwarz information criterion (BIC), for primary growth traits considered at several scales in datasets 1 and 2

Trait	Axis/scale	Dataset	Covariance structure ^a	Genotype	Tree (genotype)	Age	Year	Genotype × age	Genotype × year
L	Trunk	1	UN	**		**	**		
		2	UN	**		**	**		
	LSAS	1	UN⊗CS	**		**	**		
		2	UN⊗CS	**		**	**		
	LPAS	1	UN⊗UN	*	*	**	**		
		2	UN⊗UN	**		**	**		
	Tree	1	UN⊗UN	*	**	**	**		
		2	UN⊗UN	*	**	**	**		*
INN	Trunk	1	UN	**		**	**		
		2	UN	**		**	**		
	LSAS	1	UN⊗CS	*		**	**	*	
		2	UN⊗CS	*		**	**	*	
	LPAS	1	UN⊗UN	**		**	**		
		2	UN⊗UN	**		**	**		
	Tree	1	UN⊗UN	**	*	**	**		
		2	UN⊗UN	**	**	**	**	*	
INL	Trunk	1	ANTE(1)	**		**	**	**	
		2	UN(1)	**		**	**	**	
	LSAS	1	UN⊗CS	*	**	**	**	**	
		2	UN⊗CS	**	**	**	**	**	
	LPAS	1	UN⊗CS	**		**	**		
		2	UN⊗CS	**		**	**		
	Tree	1	UN⊗CS	**	**	**	**	**	
		2	UN⊗CS	**	**	**	**	**	
INLmax	Trunk	1	UN(1)	**		**	**		*
		2	UN(1)	**		**	**		*
	LSAS	1	UN⊗CS	**	**	**	**	*	*
		2	UN⊗UN	**	**	**	**		**
	LPAS	1	UN⊗CS	**		**	**		
		2	UN⊗CS	**		*	**		
	Tree	1	UN⊗CS	**		**	**		**
		2	UN⊗CS	**		**	**		**

^aCovariance structures: UN, an unstructured covariance matrix; UN⊗CS, a direct (Kronecker) product between an unstructured covariance matrix and a compound symmetry covariance matrix; UN⊗UN, a direct (Kronecker) product between unstructured covariance matrices; ANTE(1), a first-order antedependance covariance matrix; and UN(1), a banded main diagonal covariance matrix. In direct product between covariance matrices, the first matrix always modelled the time repetition, while the second modelled the spatial repetition. For trait abbreviations see Table 2.

internodes, they appeared to be fairly similar with few interactions between genotype and age ($G \times A$) or between genotype and year ($G \times Y$). In addition, they included a significant or highly significant effect for tree replication nested into the genotype ($T(G)$) for both traits at the whole-tree scale. Models for mean internode length and length of the longest internode included $G \times A$ and $G \times Y$ effects respectively, except on LPAS. These effects were significant or highly significant. In addition, a highly significant $T(G)$ effect was included for both traits considered on LSAS, and for mean internode length on the whole-tree scale.

Temporal repetitions were modelled by UN covariance matrices for AS length and number of internodes, and given that correlation matrices between subsequent tree ages were similar between the two traits, only those computed for AS

length are presented in Table 4. First, the matrices between datasets 1 and 2 were noticeably similar, underlining the robustness of the models used. Second, the same correlation pattern between tree ages was observed for all scales and axis types: negative correlations were observed between age 1 and subsequent ages, while positive correlations were found between subsequent ages from age 2. Covariance structures selected for mean internode length on the trunk were ANTE(1) and UN(1) in datasets 1 and 2, respectively. Whatever the scale or axis type, the correlations between ages appeared to be fairly low in comparison with those observed for other primary growth traits (Table 4). However, it was noteworthy that the correlation pattern observed for mean internode length at the whole-tree scale was close to the pattern already observed for AS length and number of internodes. UN(1) covariance

Table 4 Correlation matrices between tree ages computed from covariance matrices of models selected for two primary growth traits (L and INL) and two secondary growth traits ($\Delta Bdia$ and $Bdia$) on several tree scales in dataset 1 (lower diagonal) and dataset 2 (upper diagonal)

Primary growth																				
L									INL											
Trunk				LSAS					Trunk				LSAS							
1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	
1	1	-0.30	-0.21	-0.43	1	1	-0.22	-0.05	-0.08	1	1	-0.03	1		1	1	-0.08	0.00	-0.10	
2	-0.35	1	0.30	0.18	2	-0.20	1	0.18	0.20	2	-0.01	1		2	-0.12	1	0.11	0.11		
3	-0.19	0.36	1	0.36	3	-0.08	0.18	1	0.43	3	-0.01	0.25	1		3	0.01	0.05	1	0.14	
4	-0.41	0.22	0.37	1	4	-0.09	0.21	0.43	1	4	0.00	-0.03	-0.13	1	4	-0.10	0.09	0.12	1	
Whole tree				LPAS					Whole tree				LPAS							
1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	
1	1	-0.12	-0.16	-0.29	1					1	1	0.04	-0.05	-0.24	1					
2	-0.13	1	0.15	0.07	2			1	0.21	0.00	2	0.01	1	0.08	0.04	2		1	0.08	-0.05
3	-0.18	0.15	1	0.33	3			0.14	1	0.42	3	-0.10	0.09	1	0.15	3		0.08	1	0.29
4	-0.28	0.07	0.30	1	4			-0.11	0.31	1	4	-0.26	0.03	0.14	1	4		-0.05	0.27	1
Secondary growth																				
$\Delta Bdia$									$Bdia$											
Trunk				LSAS					Trunk				LSAS							
1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	
1	1	0.36	-0.21	-0.16	1	1	0.13	0.07	-0.04	1	1	-0.07	-0.17	-0.39	1	1	0.00	0.03	-0.23	
2	0.35	1	-0.18	-0.22	2	0.23	1	0.19	0.05	2	-0.14	1	0.36	0.18	2	0.01	1	0.33	0.13	
3	-0.01	-0.03	1	-0.20	3	0.10	0.25	1	0.00	3	-0.07	0.46	1	0.35	3	0.07	0.35	1	0.43	
4	0.00	0.01	-0.26	1	4	-0.03	0.06	0.00	1	4	-0.03	0.18	0.39	1	4	-0.18	0.13	0.41	1	
Whole tree				LPAS					Whole tree				LPAS							
1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	
1	1	0.52	0.37	0.27	1					1	1	0.17	0.03	-0.28	1					
2	0.57	1	0.28	0.18	2			1	0.10	-0.09	2	0.16	1	0.28	0.06	2		1	0.29	0.07
3	0.43	0.37	1	0.20	3			0.16	1	-0.08	3	0.06	0.29	1	0.26	3		0.25	1	0.28
4	0.28	0.20	0.22	1	4			-0.13	-0.13	1	4	-0.25	0.05	0.23	1	4		-0.03	0.15	1

Significant values ($P \leq 0.05$) according to the Wald (Z) test for covariance are indicated in bold. For trait abbreviations see Table 2.

Table 5 Significance of effects (**, highly significant, $P \leq 0.01$; *, significant, $P \leq 0.05$; ns, non significant, $P > 0.05$) in selected models according to Bayes Schwarz information criterion (BIC), for secondary growth traits considered at several scales in datasets 1 and 2

Trait	Axis/scale	Dataset	Covariance structure ^a	Genotype	Tree (genotype)	Age	Year	Genotype × age	Genotype × year
Δbdia	Trunk	1	ANTE(1)	*		**	**		
		2	UN	**		**	**		
	LSAS	1	UN⊗CS	*		**	**		
		2	UN⊗CS	**		**	**		
	LPAS	1	UN⊗UN	ns	**	**	**		
		2	UN⊗UN	ns	**	**	**		
	Tree	1	UN⊗UN	ns		**	**		*
		2	UN⊗UN	ns		**	**		**
Bdia	Trunk	1	ANTE(1)	*		**	**	*	*
		2	UN	**		**	**	**	
	LSAS	1	UN⊗CS	ns		**	**	**	
		2	UN⊗CS	*		**	**	**	
	LPAS	1	UN⊗CS	*	**	**	**		*
		2	UN⊗CS	**		**	**		
	Tree	1	UN⊗UN	ns	**	**	**	*	**
		2	UN⊗UN	*	**	**	**	*	**
Tdia	Trunk	1	ARH(1)	**		**	**		*
		2	ARH(1)	**		**	**		**
	LSAS	1	UN⊗CS	*		**	**		**
		2	UN⊗CS	**		**	**		**
	LPAS	1	UN⊗CS	ns		**	**		*
		2	UN⊗CS	**		*	**		*
	Tree	1	AR(1)⊗UN	**		**	**	*	*
		2	UN⊗UN	**		**	**	*	**

^aCovariance structures: ANTE(1), a first-order antedependance matrix; UN, an unstructured covariance matrix; UN⊗CS, a direct (Kronecker) product between an unstructured covariance matrix and a compound symmetry covariance matrix; UN⊗UN, a direct (Kronecker) product between unstructured covariance matrices; ARH(1), a heterogeneous first-order autoregressive covariance matrix; AR(1)⊗UN, a direct (Kronecker) product between a first-order autoregressive covariance matrix and an unstructured covariance matrix. In direct product between covariance matrices, the first indicated matrix always modelled the time repetition, while the second modelled the spatial repetition. For trait abbreviations see Table 2.

structures provided the best fit for length of the longest internode on the trunk, underlining the absence of any correlation between ages. In other cases, UN matrices were selected and, as already mentioned for the trunk, correlations between the subsequent ages were fairly weak (data not shown).

Spatial repetitions were modelled by a UN matrix for AS length and number of internodes, except on LSAS. For both traits, the highest correlations were found between axes belonging to the same category (i.e. between two LSAS or two LPAS), ranging from 0.41 to 0.51. But correlations were also significant between the trunk and LPAS (from 0.30 to 0.44), and between LSAS and LPAS (from 0.13 to 0.30), while correlations between the trunk and LSAS were fairly weak and not significant (< 0.1). For both mean internode length and length of the longest internode, the spatial repetitions were in all cases modelled by a CS matrix, except for length of the longest internode on LSAS in dataset 2 (UN matrix). Correlations between axes were always positive and significant, and were higher for mean internode length than for length of the longest internode (data not shown). The highest correlations

were observed between LPAS (from 0.45 to 0.52), but it is noteworthy that on the whole-tree scale the correlations between different axis types were fairly high, ranging from 0.21 to 0.25.

Secondary growth

All models built for bottom diameter increment and bottom diameter integrated G, A and Y effects (Table 5, Fig. S1b). However, even though A and Y effects were always at least significant, the significance of the G effect varied depending on trait and dataset. This was not significant for bottom diameter increment in either dataset on LPAS or on the whole-tree scale, or for bottom diameter only in dataset 1 on LSAS and on the whole-tree scale. It should be noted that few models for bottom diameter increment integrated other effects, while a $G \times A$ interaction was included for bottom diameter in all models except on LPAS. All models for top diameter included significant or highly significant G, A, Y, and $G \times Y$ effects, except on LPAS in dataset 1, where the A effect was not included and the G effect was not significant.

Covariance matrices modelling temporal repetitions for bottom diameter increment and bottom diameter were always UN, except on the trunk in dataset 1, where an ANTE(1) structure provided the best fit. For bottom diameter increment, differences in covariance structure on the trunk were related to differences in the correlation pattern (Table 4). Thus, when the trees were juvenile, correlations between ages were significant and positive in both datasets, but they became negative between subsequent ages from age 3 in dataset 2 and from age 4 in dataset 1. In other cases, correlation patterns were similar between datasets 1 and 2 for bottom diameter increment: correlations were significant and positive between all ages on the whole-tree scale, and only between subsequent ages on LSAS and LPAS. For bottom diameter, even though covariance structure differed for trunk between datasets 1 and 2, the correlation patterns were fairly similar and main differences were, rather, in the correlation values. It should be noted that this correlation pattern was similar to those previously observed for primary growth traits (negative correlations between age 1 and subsequent ages, and positive correlations between subsequent ages from age 2). Also, on other axes or on the whole-tree scale, a similar correlation pattern to those observed on the trunk was detected but only from age 2. An ARH(1) structure provided the best fit for temporal repetition when considering top diameter on the trunk. This particular structure relied on significant positive correlations between subsequent ages of *c.* 0.15 in both datasets 1 and 2 (data not shown). In other cases, the temporal repetition was modelled by a UN covariance matrix, except on the whole-tree scale in dataset 1, where an ARH(1) matrix provided the best fit. The same correlation pattern between ages was observed on LPAS and on the whole-tree scale, while on LSAS a significant positive correlation was observed only between ages 1 and 2 (data not shown).

Spatial repetition was always modelled by a UN covariance matrix for bottom diameter increment, except on LSAS, where a CS matrix provided the best fit. For bottom and top diameters, spatial repetitions were modelled by a CS matrix, except on the whole-tree scale where a UN matrix provided the best fit for both traits. For all the traits considered, correlations between axes belonging to the same category were always positive and significant (from 0.29 to 0.47). Correlations between the trunk and other axes differed between the three traits considered: for bottom diameter increment, the trunk was mostly correlated to LSAS (from 0.18 to 0.33); for bottom diameter it was mostly correlated to LPAS (from 0.20 to 0.24); and for top diameter it was significantly correlated to both LSAS and LPAS (from 0.19 to 0.45).

Branching

In all cases, the models selected for the number of axillary shoots included significant or highly significant G, A and Y effects, except on the whole-tree scale in dataset 1, where the G effect was not significant (Table 6, Fig. S1c). In both datasets,

a highly significant G \times A effect was also selected on the trunk. The models were complete on the whole-tree scale, with significant or highly significant T(G), G \times A and G \times Y effects. Considering the number of latent buds, all the models included G, A, Y and G \times Y effects. These effects were always significant or highly significant, except on LSAS where the G and A effects were not significant and were replaced by a significant G \times A effect. For the percentage of branching nodes, the models always included a highly significant G effect. On the trunk they also included significant or highly significant A, Y, and G \times Y effects, and on the whole-tree scale they were complete, with all effects significant or highly significant. By contrast, on LSAS, the models were fairly parsimonious since they included only a significant G \times A effect in addition to the genotypic effect.

For all branching traits, the temporal repetition was always modelled by a UN covariance matrix on LSAS and the whole-tree scale. On the trunk, an ANTE(1) structure provided the best fit for the number of axillary shoots. In all cases, a significant negative correlation was observed between the two first AS (from -0.23 to -0.30). Considering the number of latent buds on the trunk, a significant positive correlation was observed between the two first AS in dataset 2, while no significant correlation was detected in dataset 1. As a result, an ANTE(1) structure provided the best fit in dataset 2, while in dataset 1 a UN(1) structure was selected. In all other cases, the correlations between AS were nonsignificant. A VC structure was selected for the percentage of branching nodes on the trunk, showing that the values collected on subsequent AS were independent. On LSAS, correlations between subsequent ages were also fairly low. But, when data were gathered on the whole-tree scale, significant correlations were found between AS 1 and 2 (-0.22), and AS 1 and 3 (*c.* 0.25). These correlations might have resulted from gathering noncorrelated axes for the trait considered (see later discussion for within-tree correlations).

For all the branching traits considered, covariance matrices selected to model spatial repetitions differed depending on the scale, axis type or dataset. As branching traits were only assessed on trunk and LSAS, only the correlation between these axes could be evaluated. These were always significant between two LSAS (from 0.29 to 0.41), but correlations between trunk and LSAS differed depending on the trait considered. For the number of axillary shoots, negative correlations were computed between the trunk and LSAS (from -0.08 to -0.21). For the number of latent buds, they were positive and significant (from 0.14 to 0.21), but were fairly low and not significant for the percentage of branching nodes.

Discussion

Relevance of the staggered-start mating design for analysing genotype by environment interactions

This study aimed to dissect tree architectural plasticity into genetic, ontogenetic and environmental effects in an apple

Table 6 Significance of effects (**, highly significant, $P \leq 0.01$; *, significant, $P \leq 0.05$; ns, non significant, $P > 0.05$) in selected models according to Bayes Schwarz information criterion (BIC), for branching traits considered at several scales in datasets 1 and 2

Trait	Axis/scale	Dataset	Covariance structure ^a	Genotype	Tree (genotype)	Age	Year	Genotype × age	Genotype × year
Br	Trunk	1	ANTE(1)	*		**	**	**	
		2	ANTE(1)	*		**	**	**	
	LSAS	1	UN⊗CS	**		**	**		
		2	UN⊗CS	**		**	**		
	Tree	1	UN⊗UN	ns	**	**	**	**	*
		2	UN⊗UN	*	**	**	**	**	**
Latents	Trunk	1	UN(1)	**		**	**		*
		2	ANTE(1)	**		**	**	*	**
	LSAS	1	UN⊗UN	ns		ns	**	*	*
		2	UN⊗CS	ns		ns	**	*	**
	Tree	1	UN⊗CS	**	**	**	**		*
		2	UN⊗UN	**		**	**		**
%Br	Trunk	1	VC	**		**	**		*
		2	VC	**		**	**		**
	LSAS	1	UN⊗CS	**				*	
		2	UN⊗CS	**				*	
	Tree	1	UN⊗CS	**	**	**	**	*	**
		2	UN⊗CS	**	**	**	**	*	**

^aCovariance structures: ANTE(1), a first-order antedependance covariance matrix; UN⊗CS, a direct (Kronecker) product between an unstructured covariance matrix and a compound symmetry covariance matrix; UN⊗UN, a direct (Kronecker) product between unstructured covariance matrices; UN(1), a banded main diagonal covariance matrix; and VC, a variance components covariance matrix. In direct product between covariance matrices, the first indicated matrix always modelled the time repetition, while the second modelled the spatial repetition. For trait abbreviations see Table 2.

progeny. This was made possible through a specific mating design in which hybrids were planted 1 yr apart in the same field. Such a design, also called a staggered-start design, has been recommended in an agronomic context by Loughin (2006). Its application in a genetic context allowed us to analyze genotype × environment interactions. Multiple location designs are usually used for this purpose, but they prove costly and are not useful for phenotyping, especially when locations are very distant. As a result, few studies in perennial crops have investigated this effect, especially for architectural traits (Osorio *et al.*, 2001; Finn *et al.*, 2003; Legave *et al.*, 2006; Sykes *et al.*, 2006). In apple species, Liebhard *et al.* (2003) used a multilocation design to study certain growth traits in a 1-yr-old progeny. They highlighted interactions between genotype and environment that explained *c.* 40 and 25% of total variance for trunk height and bottom diameter, respectively. By contrast, no G × Y effect was detected for trunk length in the present study, while a significant G × Y effect was found only for trunk bottom diameter in dataset 1. The absence of any interaction for these traits in our study is likely to result from the genetic background analysed, and from the environmental conditions. In Liebhard *et al.* (2003), hybrids stemmed from another cross ('Fiesta' × 'Discovery'), and a multilocation design was used to produce marked environmental variability, including different soil conditions. By contrast, in a staggered-start design, environmental variability

is restricted to climatic conditions in subsequent calendar years. However, in the present study, a highly significant year effect was detected for most of the traits, and was interpreted as being the result of marked variability in climatic conditions. Indeed, years 2003 and 2006 were characterized by blistering summers (Fig. 2), and these extreme conditions certainly contributed to the year effect. Also, significant G × Y effects were detected for some architectural traits, such as length of the longest internode, top diameter and number of latent buds. Detecting such effects for length of the longest internode and top diameter is consistent with the fact that these traits characterize processes that occurred locally and over a short period. Indeed, the longest internode corresponds to an internode that elongates during an optimal growing period, which might differ between genotypes depending on climatic conditions. Similarly, top diameter characterizes the secondary growth that occurs at the end of the growth season over a short period. By contrast, no, or few, G × Y effects were detected for other primary and secondary growth traits characterizing processes that occur throughout the growing season. In these cases, temporary G × Y effects might have been smoothed out by integration over the entire growth season. However, this interpretation does not apply to G × Y effects detected for number of latent buds, where additional investigations would probably lead to more precise conclusions.

Within-tree variability and its consequences for phenotyping strategy

For all the traits, axes belonging to the same category (e.g. LSAS or LPAS) were always significantly correlated. From a phenotyping point of view, these results show that measures were restricted to one axis in a given category with few variability losses, as previously shown in 1-yr-old apple hybrids (Segura *et al.*, 2006). However, on the whole-tree scale the correlations between different axis categories were fairly weak for most traits. However, it is noticeable that for many architectural traits, LPAS were more correlated with the trunk than LSAS. These results are consistent with previous findings in poplars, and the assumption that sylleptic shoots are more involved in tree plasticity in response to environmental conditions than proleptic shoots (Wu & Hinckley, 2001). They also confirm marked within-tree variability, which was highlighted by differences in the model composition between axes and scales. Generally, the growth of one part of a fruit tree is considered as representative of the growth of the entire tree (Forshey & Elfving, 1989). Our results suggest that this assumption is limited to axes in the same category and underline the necessity to replicate measures on axes of different categories if the within-tree architectural variability is to be captured. In addition, particular care must be taken when selecting and gathering entities. AS length, bottom diameter, bottom diameter increment and number of axillary shoots illustrated this last point: when these variables were gathered on the whole-tree scale, the significance of the genotypic effect decreased in comparison with those estimated on separated axes.

Genotypic effect, its interaction with tree age and temporal correlations

The G effect was significant or highly significant for all primary growth traits. These results are consistent with previous inheritance studies in apple species where marked genotypic effects were shown for trunk height (Watkins & Spangelo, 1970, Liebhard *et al.*, 2003) and confirm results obtained for internode lengthening traits in this progeny over its first year of growth (Segura *et al.*, 2006, 2007). Considering primary growth variations throughout tree life (Fig. S1a), the correlation pattern observed for length and number of internodes on the trunk showed that the decline in primary growth occurred early in apple tree life, as previously found for apple cultivars (Costes *et al.*, 2003). We also demonstrated that the intensity of this decline is determined early, as revealed by the negative correlation between the first AS and subsequent years. Moreover, from the second year of growth, we detected positive correlations between subsequent AS in all axis types, including LPAS where the negative correlation between their first AS and subsequent years was no longer observed. This suggests that the effect of tree age and axis

position on the length or number of internodes for a given AS, as previously described by the concept of bud physiological age (Barthélémy *et al.*, 1997), remains consistent in a genetic context. But, since no, or few, $G \times A$ effects were detected for these two traits, the decline in the length or number of internodes with tree age appears to be homogeneous between genotypes, at least in the studied progeny. By contrast, significant $G \times A$ effects were detected for mean internode length, suggesting that this trait is differentially affected by genetic factors throughout tree life.

Considering secondary growth traits, the significance of the G effect noted for bottom diameter increment and bottom diameter varied depending on axis and scale. It was more frequently significant for top diameter than for other secondary growth traits, suggesting stronger genetic control. Considering temporal correlations, the pattern observed for trunk bottom diameter was similar to those observed for length or number of internodes (Fig. S1a,b). This is consistent with the existence of correlations between primary and secondary growth traits in a given AS, as previously found in other species (Costes *et al.*, 2000; Solar *et al.*, 2006). The correlation patterns observed for bottom diameter increment at the trunk scale differed between datasets 1 and 2, but with some similarities. In both datasets, significant positive correlations were found between the two first years of growth, and an inversion in correlation sign was observed for dataset 1 from the fourth year of growth, while in dataset 2 it occurred from the third year of growth. This difference in delay was interpreted as the consequence of fruiting occurrence, since different juvenile phase lengths were observed between datasets 1 and 2 (data not shown). This assumption is consistent with previous studies where fruiting was shown to reduce growth in apple species (see Forshey & Elfving, 1989 for a review). In particular, Visser (1970) noted a negative correlation between vigour defined as trunk bottom diameter and juvenile phase length. Also, the increase in G effect observed in dataset 2 was more pronounced than in dataset 1 for bottom diameter increment, and this increase appeared to be relatively strong in comparison to that observed on the other traits. As in a previous study, we found fairly low heritability values for secondary growth traits over the juvenile period in the same progeny (Segura *et al.*, 2006, 2007). We can thus assume that the increase in G effect for bottom diameter increment might be a consequence of its link with fruiting occurrence.

With regard to branching, the G effect was only significant for number of axillary shoots on the trunk, whereas high heritabilities were estimated when hybrids were 1 yr old (Segura *et al.*, 2006, 2007). However, at this age, only sylleptic branching was taken into account, while here the number of axillary shoots also included proleptic branching. The presence of these two shoot types on the trunk might be responsible for this decrease in G effect and may have induced a highly significant $G \times A$ effect. Conversely, on LSAS, no $G \times A$ effect

and a highly significant G effect were found, since the sylleptic process was almost anecdotic. With regard to temporal repetitions, negative correlations were found between the two first AS for number of axillary shoots in all cases, confirming the existence of ontogenetic gradients for branching as previously found in apple cultivars (Renton *et al.*, 2006). By contrast, for other branching traits, such as number of latent buds or percentage of branching nodes, the absence of any correlation between AS shows an independence of these traits over consecutive years, even though they decrease during tree ontogeny (Fig. S1c). It also suggests that the temporal correlations found for the number of axillary shoots are likely to result from those of AS length and number of internodes.

In the present study, mixed linear modelling of repeated data allowed us to dissect apple tree architectural plasticity into its genetic, ontogenetic and climatic components. The robustness of the REML method with regard to unbalancing in the dataset was confirmed for most traits by the similarity in the models between datasets 1 and 2. This validation of the model used in dataset 2 – comprising all the progeny – opens up perspectives for the identification of genomic regions involved in the architectural plasticity highlighted here. Indeed, G effects suggest the existence of stable genetic determinism with regard to tree ontogeny and climatic condition, while, by contrast, $G \times A$, $G \times Y$, and within-tree variability suggest specific genetic determinism for ontogenetic and climatic effects. As a result, a QTL analysis based on the best linear unbiased predictors (BLUP) for G, $G \times A$, and $G \times Y$ effects has been initiated to complement the results presented here.

Acknowledgements

We are grateful to Gilbert Garcia and Stephan Feral for their contribution to field measurements and their technical assistance in the orchard. We also acknowledge Mark Jones for improving the English, and two anonymous referees for helpful comments on the manuscript. This research was partly funded by a grant from the INRA genetic and plant breeding department and Languedoc-Roussillon region, allocated to Vincent Segura.

References

- Barthélémy D, Caraglio Y, Costes E. 1997. Architecture, gradients morphogénétiques et âge physiologique chez les végétaux. In: Bouchon J, Reffye de P, Barthelemy D, eds. *Modélisation et simulation de l'architecture des végétaux*. Paris, France: INRA éditions, 89–136.
- Bond BJ. 2000. Age-related changes in photosynthesis in woody plants. *Trends in Plant Science* 5: 349–353.
- Bond BJ, Czarnomski NM, Cooper C, Day ME, Greenwood M.S. 2007. Developmental decline in height growth in Douglas-fir. *Tree Physiology* 27: 441–453.
- Bradshaw HD Jr, Stettler RF. 1995. Molecular genetics of growth and development in populus. IV. mapping QTLs with large effects on growth, form, and phenology traits in a forest tree. *Genetics* 139: 963–973.
- Comstock RE, Moll RH. 1963. Genotype-environment interactions. In: Hanson WD, Robinson HF, eds. *Statistical genetics and plant breeding*. Washington, DC, USA: National Academy of Sciences, 164–196.
- Conner PJ, Brown SK, Weeden NF. 1998. Molecular-marker analysis of quantitative traits for growth and development in juvenile apple trees. *Theoretical and Applied Genetics* 96: 1027–1035.
- Costes E, Fournier D, Salles JC. 2000. Changes in primary and secondary growth as influenced by crop load in 'Fantasme'® apricot trees. *The Journal of Horticultural Science and Biotechnology* 75: 510–519.
- Costes E, Guédon Y. 2002. Modelling branching patterns on one-year-old trunks of apple cultivars. *Annals of Botany (Lond)* 89: 513–523.
- Costes E, Lauri PE, Regnard JL. 2006. Analysing fruit tree architecture, implication for tree management and fruit production. *Horticultural Reviews* 32: 1–61.
- Costes E, Sinoquet H, Kelner JJ, Godin C. 2003. Exploring within-tree architectural development of two apple tree cultivars over 6 years. *Annals of Botany (Lond)* 91: 91–104.
- Dieters MJ, White TL, Littell RC, Hedge GR. 1995. Application of approximate variances of variance-components and their ratios in genetic tests. *Theoretical and Applied Genetics* 91: 15–24.
- Finn CE, Hancock JF, Mackey T, Serce S. 2003. Genotype \times environment interactions in highbush blueberry (*Vaccinium* sp. L.) families grown in Michigan and Oregon. *Journal of the American Society for Horticultural Science* 128: 196–200.
- Forshey CG, Elfving DC. 1989. The relationship between vegetative growth and fruiting in apple trees. *Horticultural Reviews* 11: 229–287.
- Holland JB. 2007. Genetic architecture of complex traits in plants. *Current Opinion in Plant Biology* 10: 156–161.
- Kenis K, Keulemans J. 2007. Study of tree architecture of apple (*Malus \times domestica* Borkh.) by QTL analysis of growth traits. *Molecular Breeding* 19: 193–208.
- Laurens F, Audergon J, Claverie J, Duval H, Germain E, Kervella J, Lelezec M, Lauri P, Lespinasse JM. 2000. Integration of architectural types in French programmes of ligneous fruit species genetic improvement. *Fruits* 55: 141–152.
- Legave JM, Segura V, Fournier D, Costes E. 2006. The effect of genotype, location and their interaction on early growth and branching in apricot trees. *The Journal of Horticultural Science and Biotechnology* 81: 189–198.
- Lespinasse JM. 1977. *La conduite du pommier: types de fructification, incidence sur la conduite de l'arbre*, Vol. 1. Paris, France: INVUFLEC.
- Lespinasse Y. 1992. Le pommier. In: Gallais A, Bannierot H, eds. *Amélioration des espèces végétales cultivées – objectifs et critères de sélection*. Paris, France: INRA éditions, 579–594.
- Liebhart R, Kellerhals M, Pfammatter W, Jertmini M, Gessler C. 2003. Mapping quantitative physiological traits in apple (*Malus \times domestica* Borkh.). *Plant Molecular Biology* 52: 511–526.
- Littell RC, Henry PR, Ammerman CB. 1998. Statistical analysis of repeated measures data using SAS procedures. *Journal of Animal Science* 76: 1216–1231.
- Littell RC, Pendergast J, Natarajan R. 2000. Modelling covariance structure in the analysis of repeated measures data. *Statistics in Medicine* 19: 1793–1819.
- Loughin TM. 2006. Improved experimental design and analysis for long-term experiments. *Crop Science* 46: 2492–2502.
- Lynch M, Walsh B. 1998. *Genetics and analysis of quantitative traits*. Sunderland, MA, USA: Sinauer Associates.
- Moser BE. 2004. Repeated measures modeling with PROC MIXED. 29th Annual SAS Users Group International Conference, Montréal, Canada. Paper 188–129. [<http://www2.sas.com/proceedings/sugi29/188-29.pdf>].
- Osorio LF, White TL, Huber DA. 2001. Age trends of heritabilities and genotype-by-environment interactions for growth traits and wood density from clonal trials of *Eucalyptus grandis* Hill ex Maiden. *Silvae Genetica* 50: 30–37.

- Plomion C, Durel CE, O'Malley DM. 1996. Genetic dissection of height in maritime pine seedlings raised under accelerated growth conditions. *Theoretical and Applied Genetics* 93: 849–858.
- Renton M, Guédon Y, Godin C, Costes E. 2006. Similarities and gradients in growth unit branching patterns during ontogeny in 'Fuji' apple trees: a stochastic approach. *Journal of Experimental Botany* 57: 3131–3143.
- SAS Institute Inc. 2000. *SAS user's guide: statistics*. Cary, NC, USA: SAS Institute Inc.
- Scotti-Saintagne C, Bodénès C, Barreneche T, Bertocchi E, Plomion C, Kremer A. 2004. Detection of quantitative trait loci controlling bud burst and height growth in *Quercus robur* L. *Theoretical and Applied Genetics* 109: 1648–1659.
- Segura V, Cilas C, Laurens F, Costes E. 2006. Phenotyping progenies for complex architectural traits: a strategy for 1-year-old apple trees (*Malus × domestica* Borkh.). *Tree Genetics and Genomes* 2: 140–151.
- Segura V, Denancé C, Durel CE, Costes E. 2007. Wide range QTL analysis for complex architectural traits in a 1-year-old apple progeny. *Genome* 50: 159–171.
- Seleznyova AN, Thorp T, White M, Tustin S, Costes E. 2003. Application of architectural analysis and AMAPmod methodology to study dwarfing phenomenon: the branch structure of 'Royal Gala' apple grafted on dwarfing and non-dwarfing rootstock combinations. *Annals of Botany (Lond)* 91: 1–8.
- Solar A, Solar M, Štampar F. 2006. Stability of the annual shoot diameter in Persian walnut: a case study of different morphotypes and years. *Trees* 20: 449–459.
- Sykes R, Li B, Isik F, Kadla J, Chang H-M. 2006. Genetic variation and genotype by environment interactions of juvenile wood chemical properties in *Pinus taeda* L. *Annals of Forest Science* 63: 897–904.
- Visser T. 1970. The relation between growth, juvenile period and fruiting of apple seedlings and its use to improve breeding efficiency. *Euphytica* 19: 293–302.
- Watkins R, Spangelo LPS. 1970. Components of genetic variance for plant survival and vigor of apple trees. *Theoretical and Applied Genetics* 40: 195–203.
- White J. 1979. The plant as a metapopulation. *Annual Review of Ecology and Systematics* 10: 109–145.
- Wu RL. 1998. Genetic mapping of QTLs affecting tree growth and architecture in *Populus*: implication for ideotype breeding. *Theoretical and Applied Genetics* 96: 447–457.
- Wu R, Hinckley TM. 2001. Phenotypic plasticity of sylleptic branching: genetic design of tree architecture. *Critical Reviews in Plant Science* 20: 467–485.
- Wu RL, Lin M. 2006. Functional mapping – how to map and study the genetic architecture of dynamic complex traits. *Nature Reviews Genetics* 7: 229–237.
- Wu R, Ma C-X, Lin M, Casella G. 2004. A general framework for analyzing the genetic architecture of developmental characteristics. *Genetics* 166: 1541–1551.
- Wu R, Ma C-X, Zhao W, Casella G. 2003. Functional mapping for quantitative trait loci governing growth rates: a parametric model. *Physiological Genomics* 14: 241–249.
- Wu R, Stettler RF. 1996. The genetic resolution of juvenile canopy structure and function in a three-generation pedigree of *Populus*. *Trees – Structure and Function* 11: 99–108.
- Wu R, Stettler RF. 1998. Quantitative genetics of growth and development in *Populus*. III. Phenotypic plasticity of crown structure and function. *Heredity* 81: 299–310.
- Yang R, Tian Q, Xu S. 2006. Mapping quantitative trait loci for longitudinal traits in line crosses. *Genetics* 173: 2339–2356.

Supplementary Material

The following supplementary material is available for this article online:

Fig. S1 Changes in architectural trait means throughout the first 4 yr of growth. Data from the 2003 design (squares), from the 2004 design (triangles), and all data (circles). In this latter case, standard deviations are also indicated. Traits are presented by architectural process: (a) primary growth, (b) secondary growth, and (c) branching. For trait abbreviations, see Table 2.

This material is available as part of the online article from: <http://www.blackwell-synergy.com/doi/abs/10.1111/j.1469-8137.2008.02374.x> (This link will take you to the article abstract).

Please note: Blackwell Publishing are not responsible for the content or functionality of any supplementary materials supplied by the authors. Any queries (other than missing material) should be directed to the journal at *New Phytologist* Central Office.

The Genetic Map of *Artemisia annua* L. Identifies Loci Affecting Yield of the Antimalarial Drug Artemisinin

Ian A. Graham,^{1*} Katrin Besser,¹ Susan Blumer,¹ Caroline A. Branigan,¹ Tomasz Czechowski,¹ Luisa Elias,¹ Inna Guterman,¹ David Harvey,¹ Peter G. Isaac,² Awais M. Khan,¹ Tony R. Larson,¹ Yi Li,¹ Tanya Pawson,¹ Teresa Penfield,¹ Anne M. Rae,¹ Deborah A. Rathbone,¹ Sonja Reid,¹ Joe Ross,¹ Margaret F. Smallwood,¹ Vincent Segura,¹ Theresa Townsend,¹ Darshna Vyas,¹ Thilo Winzer,¹ Dianna Bowles^{1*}

Artemisinin is a plant natural product produced by *Artemisia annua* and the active ingredient in the most effective treatment for malaria. Efforts to eradicate malaria are increasing demand for an affordable, high-quality, robust supply of artemisinin. We performed deep sequencing on the transcriptome of *A. annua* to identify genes and markers for fast-track breeding. Extensive genetic variation enabled us to build a detailed genetic map with nine linkage groups. Replicated field trials resulted in a quantitative trait loci (QTL) map that accounts for a significant amount of the variation in key traits controlling artemisinin yield. Enrichment for positive QTLs in parents of new high-yielding hybrids confirms that the knowledge and tools to convert *A. annua* into a robust crop are now available.

Malaria is a global health problem with more than 1 billion people living in areas with a high risk of the disease. Artemisinin combination therapies (ACTs) are the recommended treatment for uncomplicated malaria caused by the *Plasmodium falciparum* parasite (1). Parasite resistance to artemisinin has recently been confirmed in western Cambodia (2). It has long been recognized that the problem of artemisinin resistance is best addressed by increasing access to ACTs and discouraging the use of artemisinin monotherapies (3). This approach has strong support from the global health community with both funding and demand for ACTs expected to increase massively in the short- to midterm (3). However, there is growing concern that the supply chain will be unable to consistently produce high-quality artemisinin in the quantities that will be required (3). Artemisinin is a sesquiterpenoid synthesized in the glandular trichomes of the Chinese medicinal plant *Artemisia annua* L. (4–10). For a pharmaceutical with annual sales exceeding 100 million treatments, ACT supply remains reliant on the agricultural production of artemisinin. Plant-based production of artemisinin is challenging because *A. annua* remains relatively undeveloped as a crop. An alternative microbial-based system that synthesizes an artemisinin precursor for chemical conversion is in development (11, 12). This would supplement but not replace agricultural production, which will continue to be an essential source of supply (3). Improved varieties of *A. annua* for developing-world farmers would bring immediate benefits to the existing artemisinin supply chain by reducing production costs, stabilizing supplies, and improving grower confidence in the crop (3).

¹Centre for Novel Agricultural Products, Department of Biology, University of York, York YO10 5YW, UK. ²IDna Genetics Ltd., Norwich Research Park, Norwich NR4 7UH, UK.

*To whom correspondence should be addressed. E-mail: iag1@york.ac.uk (I.A.G.); djb32@york.ac.uk (D.B.)

A. annua is a member of the *Asteraceae* family that favors outcrossing over selfing (13). The artemisinin content of plants from different origins varies considerably and is highly heritable (14). The market leader for artemisinin production, at present, is Artemis, an F₁ hybrid (population) variety developed by Mediplant (Conthey, Switzerland)

(14). Artemis seed is produced from a cross between two heterozygous and genetically different parental genotypes, called C4 and C1, that are themselves maintained vegetatively. In this study, we have used the Artemis pedigree to establish genetic linkage and QTL maps for this species and independently validated positive QTL for artemisinin yield.

We used the Roche 454 pyrosequencing platform to produce expressed sequence tag (EST) databases from cDNA libraries derived from enriched glandular trichome preparations of young leaves, mature leaves, and flower buds from the Artemis hybrid (15). cDNA libraries and EST databases were also prepared from meristem tissue (including very young leaf tissue) and cotyledons. A selection of key genes associated with metabolic pathways and phenotypic traits such as trichome development and plant architecture that could affect artemisinin yield are illustrated in fig. S1, together with their relative abundance in the different libraries (SOM Text and table S1). The EST sequences were also used for in silico identification of single-nucleotide polymorphisms (SNPs), short sequence repeats (SSRs), and insertions/deletions (InDels), which can be used as molecular markers for mapping and breeding (15). We identified 34,419 SNPs from DNA sequences contained in the five EST databases derived from the Artemis F₁ hybrid material, representing a mean SNP frequency of 1 in 104 base pairs

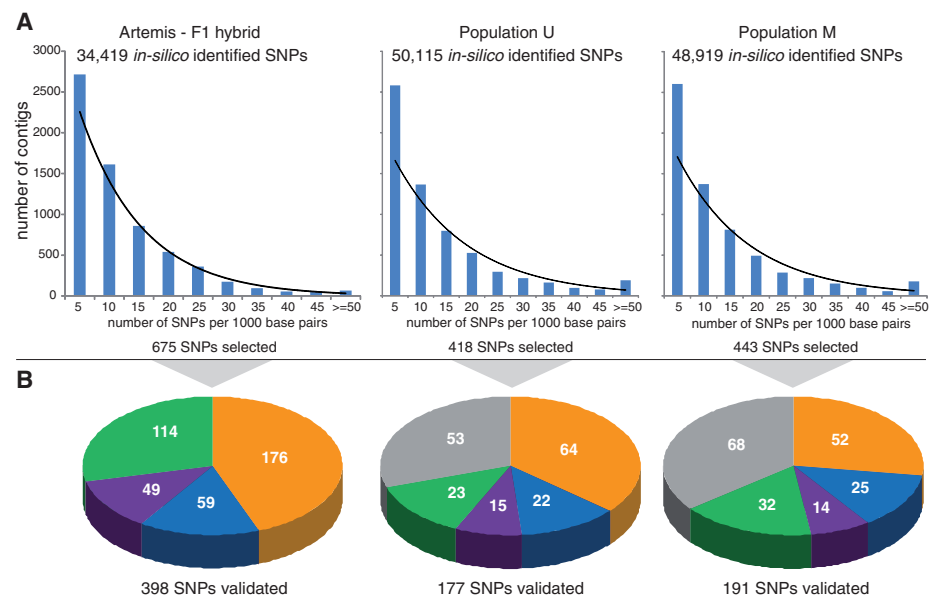


Fig. 1. High-throughput identification and validation of SNP markers in three *A. annua* populations. **(A)** Frequency distribution of potential SNPs identified in silico from EST databases produced by pyrosequencing cDNA libraries from the Artemis F₁ hybrid, Population U (commercially grown in Uganda) and Population M (commercially grown in Madagascar). The observed distribution of SNP frequency correlates closely with an exponential distribution for each data set as indicated by the curved black lines, which trace the expected distribution and R^2 values that are greater than 0.85 in all cases. Stringent selection criteria resulted in approximately only 10% of contigs being used for SNP identification (15). **(B)** Genotyping the Artemis pedigree. Subsets of in silico-identified SNPs from each population were selected for hybridization-based detection on the Illumina Goldengate Genotyping platform. The three pie charts show the genotyping of the Artemis pedigree with SNPs from Artemis, and Populations U and M. Color coding illustrates the proportion of SNPs that were polymorphic in the C4 parent (orange), polymorphic in the C1 parent (blue), polymorphic in both parents (purple), monomorphic in both parents for opposite alleles (green), and monomorphic in both parents for the same alleles (gray). This latter class is due to alleles being polymorphic in Population M or Population U but not in Artemis.

(Fig. 1A). This polymorphism was confirmed experimentally with 19 amplified fragment length polymorphism (AFLP) primer combinations that revealed 322 polymorphic markers (table S2). The

in silico approach also identified 49 SSR markers that segregated in the Artemis F₁ population (table S3). We extended the in silico approach to two other *A. annua* populations, commercially grown

in Uganda (Population U) and Madagascar (Population M), and found that the mean SNP frequencies of 1 in 88 and 1 in 91 base pairs, respectively, are only slightly higher than that of the Artemis F₁ hybrid (Fig. 1A).

We used the Illumina GoldenGate Genotyping platform to exploit this genetic resource, employing stringent criteria for selection of 1536 SNPs from the pool of 133,000 in silico-identified SNPs from Artemis and Populations U and M (15). The subset of SNPs represented candidate genes and their homologs, as well as others chosen randomly with the aim of having well-spaced markers for the genetic linkage map. We developed size-based markers in addition for 104 of the 1536 SNPs that allowed the two alleles in each case to be distinguished by capillary electrophoresis, and these further confirmed the segregation data derived from the Illumina platform (table S3). Genotyping the Artemis pedigree confirmed that extensive heterozygosity exists in the Artemis parents (Fig. 1B). Of SNPs derived from Populations U and M, 70% and 64%, respectively, were also found to segregate in the Artemis pedigree. The heterozygosity in C4 is roughly double that of C1, reflecting differences in the history of these genotypes. A number of markers are monomorphic in both parents for opposite alleles. These fixed differences between parents will segregate in generations beyond the F₁, thereby offering additional segregation of alleles not revealed in Artemis.

Phenotypic variation can be seen in the Artemis pedigree, consistent with the high level of genetic variation. This is shown in Fig. 2 for our mapping population of the Artemis F₁, grown in UK field trials during 2007 (UK07). Metabolite profiling revealed concentrations of artemisinin that ranged from 0.93 to 20.65 μg/mg dry weight, with associated metabolites also showing variation (Fig. 2A). Leaf area ranged between 508.76 and 4696.08 mm² (Fig. 2B), glandular trichome density between 4.89 and 19.11 mm⁻² (Fig. 2C), and plant fresh weight between 160 and 4440 g (Fig. 2D). These traits are targets for increasing artemisinin yield, which is a product of both artemisinin concentration and plant fresh weight.

The fact that the Artemis parents are heterozygous enabled us to produce genetic linkage maps for each parent based on data derived from an F₁ mapping population of 242 individuals (fig. S2) (15). Using a minimum LOD (logarithm of the odds ratio for linkage) score of 4.0, we defined nine linkage groups for the C4 parent and seven linkage groups for C1 (fig. S2). We hypothesized that the C1 map is missing two linkage groups, designated LG8 and LG9, because two chromosomes in the C1 parent are either homozygous or have a very low level of heterozygosity and therefore do not segregate for markers from this parent in the F₁, so cannot be mapped in this generation. To test this hypothesis, an individual F₁ plant showing high heterozygosity in molecular analysis was self-pollinated to produce an F₂ generation. Markers seen to be homozygous for opposite alleles in the parents, and therefore heterozygous in all F₁ progeny, were genotyped in the F₂ generation

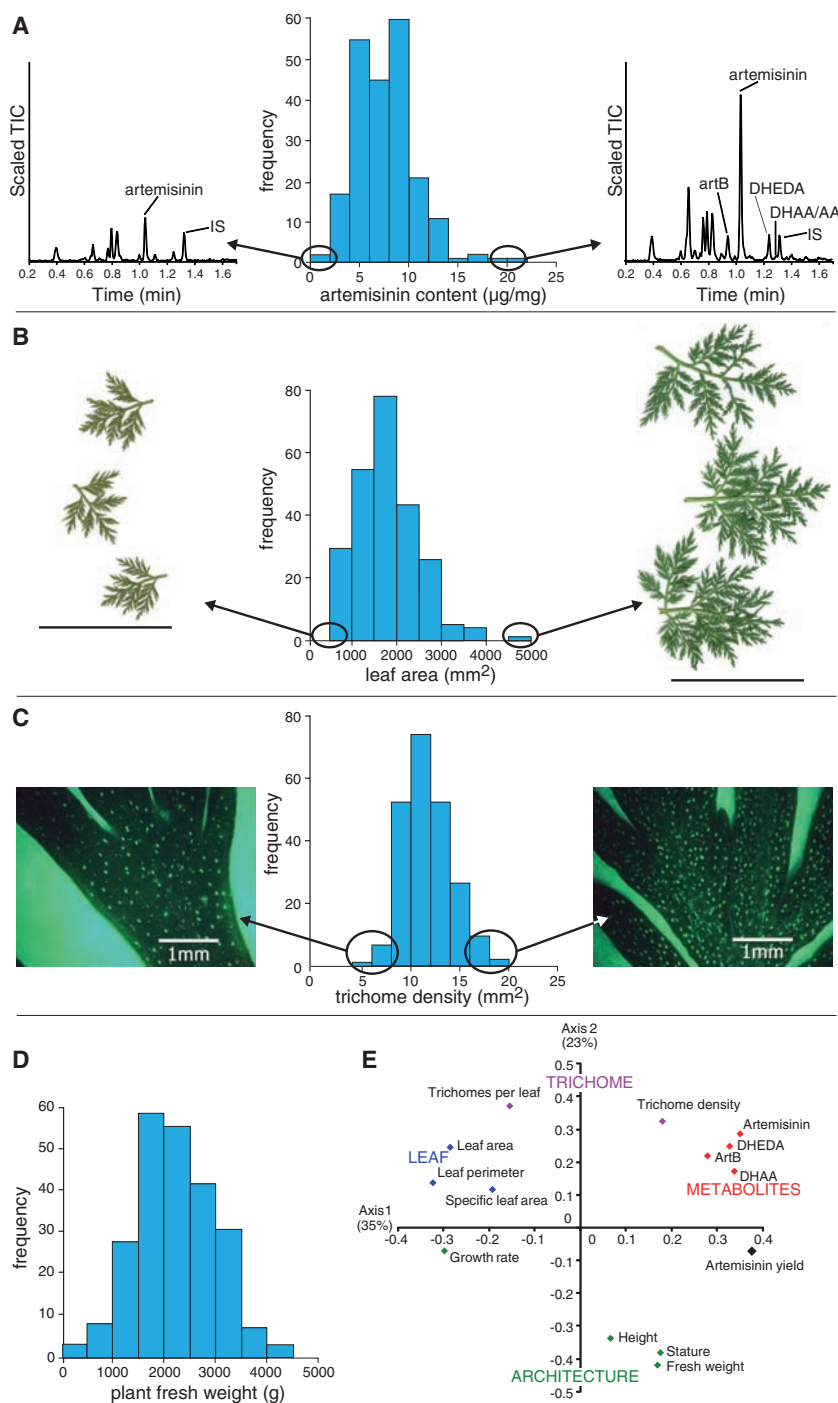


Fig. 2. Phenotypic variation in the Artemis F₁ grown in UK field trials in 2007. The distribution of four traits related to artemisinin yield is illustrated. (A) Artemisinin concentration at harvest (7 months after sowing). Metabolite profiles showing artemisinin and related metabolites from the lowest- and highest-yielding plants relative to an internal standard (IS) are presented. artB, arteannuin B; DHEDA, dihydro-epi-deoxyarteannuin B; DHAA, dihydroartemisinic acid; AA, artemisinic acid. (B) Leaf area 5 months after sowing. Images show leaves from positions 20, 21, and 22 from the apical meristem. (C) Trichome density 5 months after sowing. The abaxial surface of leaves 15, 16, and 17 from the apical meristem was visualized by fluorescent microscopy. Glandular trichomes appear as bright green spots. (D) Fresh weight of aboveground plant material at harvest (7 months after sowing). (E) Principal component analysis of traits related to artemisinin yield. Architecture, leaf, and metabolite traits additional to those detailed in (A) to (D) are also included in the analysis as shown.

together with markers known to map to LG8 and LG9 in the C4 parent. In support of our hypothesis, a number of these markers were found to segregate in the F₂ generation. These data allowed F₂ linkage groups for LG8 and LG9 to be defined and anchored to the corresponding C4 linkage groups (fig. S2). The identification of nine LGs is consistent with cytological studies reporting the diploid number of chromosomes to be 18 in *A. annua* (16). The marker positions shown on the map were validated by three

independent approaches: coalignment on the C4 and C1 maps, common location of multiple markers from single candidate genes, and robustness of marker order after reconstruction of the map with a subset of markers (SOM Text).

We used vegetative propagation to replicate individuals from the mapping population, which enabled us to perform three independent field trials using the same genotypes. A single replicate of each genotype was tested in 2007 in the UK

(UK07) and three replicates of each genotype were tested both in the UK (UK08) and Switzerland (SW08) in 2008. Fourteen traits were scored that could affect artemisinin yield (Fig. 2E). All these traits exhibited a moderate to high heritability ranging from 0.41 to 0.62, resulting in the discovery of multiple QTLs. Stable QTLs for artemisinin concentration were identified on C4 LG1, LG4, and LG9 (Fig. 3 and table S4), which describe 20% of the variation in UK07 and between 30 and 38% in

Fig. 3. A selection of QTLs for key traits identified across three field trials. QTLs are shown to the right and distances in centimorgans to the left of each linkage group. Thick and thin lines indicate the confidence intervals of the QTLs corresponding to 1 and 2 LOD units below the maximum LOD score, respectively. QTLs are shown for artemisinin concentration (in red), artemisinin yield (artemisinin concentration × fresh weight) (in black), architecture (fresh weight and stature) (in green), and leaf area (in blue). Trials in which QTLs were detected are denoted as UK07, UK08 and SW08. Candidate genes associated with QTL are *DXR2* (1-Deoxy-D-Xylulose 5-phosphate Reductoisomerase 2) and *MAX3* (More Axillary Branching 3).

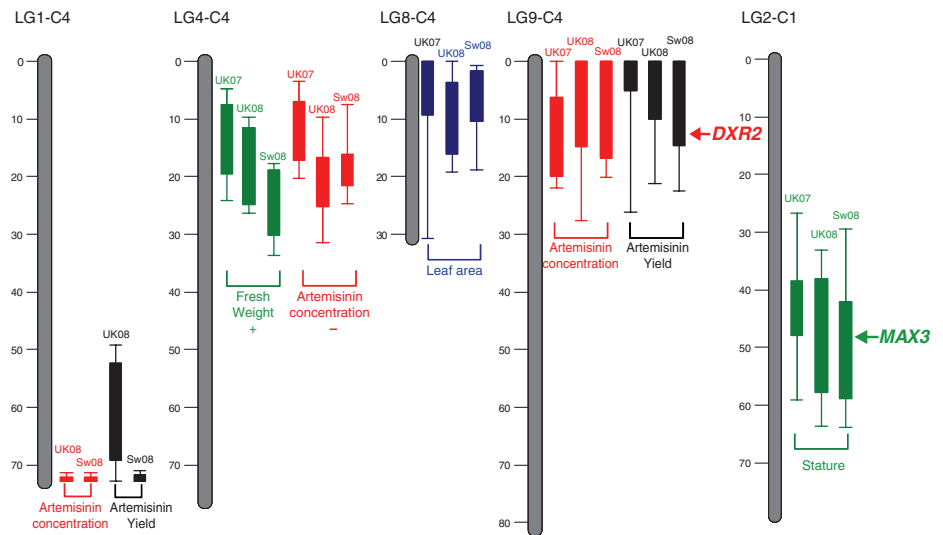
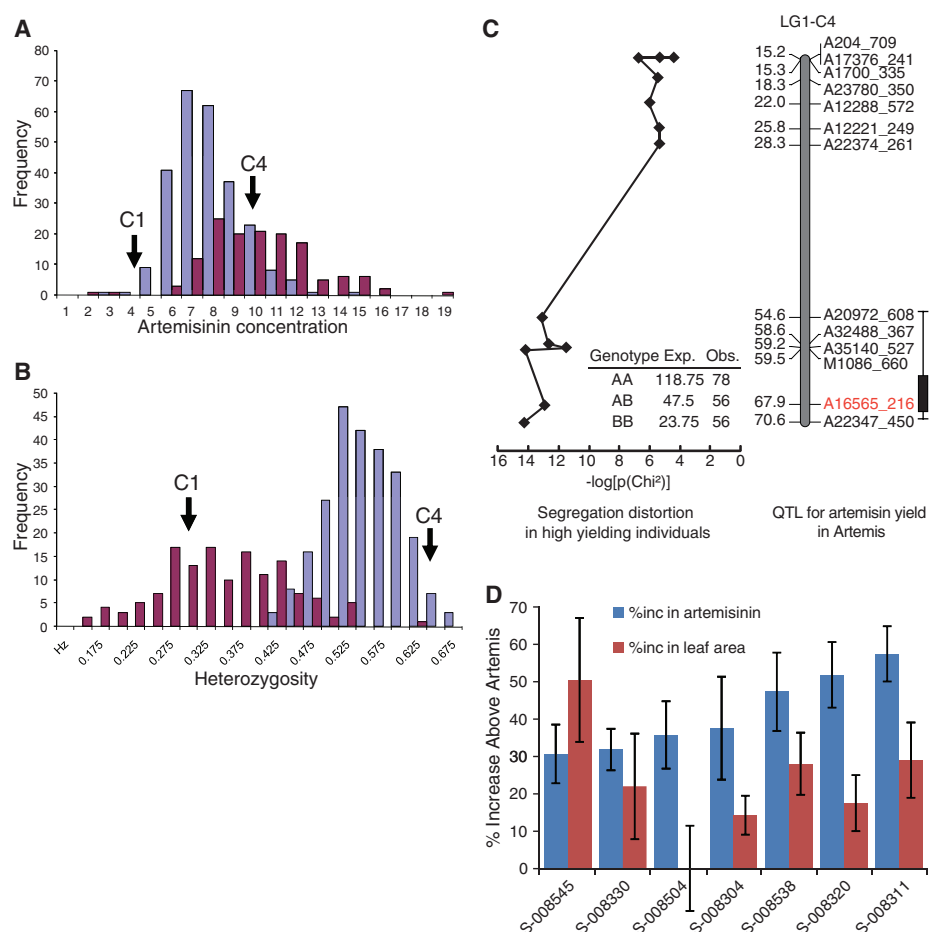


Fig. 4. Genetic analysis of high-yielding plants. (A) Distribution of artemisinin concentration (μg/mg dry weight) in the F₁ mapping population of 242 individuals is shown in blue and that in 130 selected high-yielding F₂ individuals grown in the same trial (UK08), is shown in red. The artemisinin concentrations of C4 and C1 grown in the same trial are indicated. (B) Distribution of heterozygosity scores for the same individuals as in (A). (C) The position of a major QTL for artemisinin yield on LG1 and markers in this region show high segregation distortion in favor of the increasing alleles in 190 F₂ high-yielding individuals. For the marker highlighted in red, the B allele has a positive effect on yield ($P = 4.4 \times 10^{-7}$) and is overrepresented in the high-yielding individuals summarized in the table inset. The plotted values for segregation distortion represent the $-\log[\text{Chi-squared}]$ based on the observed and expected values for genotype classes at a number of markers on linkage group 1. (D) The percentage increase in artemisinin concentration (in red) and leaf area (in blue), over Artemis F₁ for seven hybrids produced from crosses of selected high-yielding individuals. Values are the mean ± SE for a minimum of five individual replicates.



UK08 and SW08. Artemisinin yield is a product of both artemisinin concentration and fresh weight. QTLs for yield collocate to those for artemisinin concentration on LG1 and LG9, thus representing targets for a breeding program. The artemisinin concentration QTL on LG4 collocates with a QTL for fresh weight but with opposing effects on artemisinin yield (Fig. 3). Markers in candidate genes collocate with a number of the QTLs. For example, the precursor supply gene candidate *DXR2* collocates with the QTL for artemisinin yield on C4 LG9 and the architectural trait candidate gene *MAX3* collocates with the QTL for stature on C1 LG2.

In parallel with the development of the marker-assisted breeding program, we performed a high-throughput screen for artemisinin content in 23,000 12-week-old glasshouse-grown F₂ and F₃ plants derived from F₁ Artemis seed that had been mutagenized with ethylmethane sulfonate (15). The mutation frequency in this material was determined with the TILLING method and found to be approximately one EMS-induced mutation per 5.4 Mb (15). This is less than the SNP frequency determined for Artemis at one polymorphism per 104 base pairs. This screen should therefore identify individuals carrying beneficial mutations derived from the EMS treatment and also individuals carrying improved genetic backgrounds as a result of segregation of favorable alleles derived from natural variation. We found that the distribution of artemisinin content among selected high-yielding F₂ individuals is higher than in the UK08 Artemis F₁ mapping population (Fig. 4A) even though overall heterozygosity is lower (Fig. 4B). Next, we determined whether any of the QTLs we had identified for artemisinin yield on the basis of field trials are overrepresented in the high-yielding

individuals that had been selected under glasshouse conditions. We found strong segregation distortion in favor of the advantageous alleles for an artemisinin yield QTL on C4 LG1 (Fig. 4C). These data validate this QTL and confirm that for artemisinin yield, the genotype has a strong influence on both glasshouse-grown and field-grown material.

An ongoing empirical hybridization program of high-yielding plants identified in the high-throughput phenotypic screen produced hybrid progeny that outperformed Artemis for artemisinin concentration and leaf area after 12 weeks' growth under glass (Fig. 4D). The choice of parents for this program preceded the availability of QTL data and was based on phenotypic characteristics (15). In terms of utility in a molecular breeding program, we found a significant association of positive artemisinin yield QTL in those parents that produced hybrids with increased artemisinin yield ($P < 0.001$).

Our study has established the molecular basis for marker-assisted breeding of this medicinal plant species and highlights the reduced timelines that are now feasible for developing this platform of knowledge and tools. The artemisinin from *A. annua* is the key component in the ACT treatment of malaria, and demand for ACTs is expected to increase in the immediate future. Development of new high-yielding varieties optimized for production in different geographic regions is now a realistic target.

References and Notes

1. World Malaria Report 2008, World Health Organisation; <http://apps.who.int/malaria/wmr2008/malaria2008.pdf>.
2. A. M. Dondorp et al., *N. Engl. J. Med.* **361**, 455 (2009).
3. "Saving Lives, Buying Time: Economics of Malaria Drugs in an Age of Resistance," National Academy of Sciences 2004, www.nap.edu/catalog/11017.html. Global Malaria Action Plan, Report of the 2008 Artemisinin

- Conference, 8 to 10 October, York, UK (www.york.ac.uk/org/cnap/artemisiaproject/pdfs/AEconference-report-web.pdf).
4. M. V. Duke, R. N. Paul, H. N. Elsohly, G. Sturtz, S. O. Duke, *Int. J. Plant Sci.* **155**, 365 (1994).
5. C. M. Berteau et al., *Planta Med.* **71**, 40 (2005).
6. C. M. Berteau et al., *Arch. Biochem. Biophys.* **448**, 3 (2006).
7. K. H. Teoh, D. R. Polichuk, D. W. Reed, G. Nowak, P. S. Covello, *FEBS Lett.* **580**, 1411 (2006).
8. Y. Zhang et al., *J. Biol. Chem.* **283**, 21501 (2008).
9. K. H. Teoh, D. R. Polichuk, D. W. Reed, P. S. Covello, *Can. J. Bot.* **87**, 635 (2009).
10. P. S. Covello, *Phytochemistry* **69**, 2881 (2008).
11. M. C. Chang, R. A. Eachus, W. Trieu, D. K. Ro, J. D. Keasling, *Nat. Chem. Biol.* **3**, 274 (2007).
12. D. K. Ro et al., *Nature* **440**, 940 (2006).
13. J. F. S. Ferreira, J. Janick, *Int. J. Plant Sci.* **156**, 807 (1995).
14. N. Delabays, X. Simonnet, M. Gaudin, *Curr. Med. Chem.* **8**, 1795 (2001).
15. Information on materials and methods is available on Science Online.
16. M. Torrell, J. Vallés, *Genome* **44**, 231 (2001).
17. We thank L. Doucet, H. Martin, N. Nattriss, M. Segura, and A. Czechowska for horticulture assistance; G. Chigeza for horticulture management; S. Graham, S. Heywood, B. Kowalik, S. Pandey, R. Simister, and C. Whitehead for laboratory assistance; C. Calvert, P. Dicks, W. Lawley, and D. Rotherham for project management; E. Bartlett for communications advice; and P. Roberts for graphic design. We thank L. Brewer, H. Klee, and K. Stuart for insightful advice on this project. We thank X. Simonnet and Médiplant for access to the Artemis pedigree. We acknowledge financial support for this project from The Bill and Melinda Gates Foundation and Medicines for Malaria Venture, as well as from The Garfield Weston Foundation for the Centre for Novel Agricultural Products.

Supporting Online Material

www.sciencemag.org/cgi/content/full/327/5963/328/DC1
Materials and Methods
SOM Text
Figs. S1 to S6
Tables S1 to S4
References

29 September 2009; accepted 20 November 2009
10.1126/science.1182612

Tetrathiomolybdate Inhibits Copper Trafficking Proteins Through Metal Cluster Formation

Hamsell M. Alvarez,^{1*} Yi Xue,^{1*} Chandler D. Robinson,¹ Mónica A. Canalizo-Hernández,¹ Rebecca G. Marvin,¹ Rebekah A. Kelly,³ Alfonso Mondragón,² James E. Penner-Hahn,³ Thomas V. O'Halloran^{1,2†}

Tetrathiomolybdate (TM) is an orally active agent for treatment of disorders of copper metabolism. Here we describe how TM inhibits proteins that regulate copper physiology. Crystallographic results reveal that the surprising stability of the drug complex with the metallochaperone Atx1 arises from formation of a sulfur-bridged copper-molybdenum cluster reminiscent of those found in molybdenum and iron sulfur proteins. Spectroscopic studies indicate that this cluster is stable in solution and corresponds to physiological clusters isolated from TM-treated Wilson's disease animal models. Finally, mechanistic studies show that the drug-metallochaperone inhibits metal transfer functions between copper-trafficking proteins. The results are consistent with a model wherein TM can directly and reversibly down-regulate copper delivery to secreted metalloenzymes and suggest that proteins involved in metal regulation might be fruitful drug targets.

Excess dietary molybdate (MoO_4^{2-}) uptake was first linked to a fatal disorder in cattle known as "teart" pastures syndrome (1) and later to a neurological disorder in sheep

known as "swayback" (2). Both disorders arise from Mo-induced copper deficiency, and the symptoms are readily reversed with copper supplementation. Although molybdate itself has little or no

affinity for copper ions, the active copper-depleting agent, TM (MoS_4^{2-}), is formed in the ruminants' digestive track and readily reacts with Cu^{I} or Cu^{II} to form insoluble compounds. These zoogenic studies inspired the development of molybdenum compounds to treat copper-dependent diseases in humans (3). The potent chelating and antiangiogenic activities of orally active formulations of TM, such as the ammonium salt $[(\text{NH}_4)_2(\text{MoS}_4)]$ (4–6) and the choline salt (ATN-224) (7, 8), have been used in treatment of Wilson's disease, where copper accumulation leads to hepatic and neurological disorders, as well as in the inhibition of metastatic cancer progression in a number of clinical trials (9–11). TM inhibits several copper enzymes, including ceruloplasmin (Cp), ascorbate oxidase, cytochrome oxidase, superoxide dismutase (SOD1), tyrosinase, and the *Enterococcus*

¹The Chemistry of Life Processes Institute, Northwestern University, Evanston, IL 60208, USA. ²Department of Biochemistry, Molecular Biology and Cell Biology, Northwestern University, Evanston, IL 60208, USA. ³Department of Chemistry, The University of Michigan, Ann Arbor, MI 48109, USA.

*These authors contributed equally to this work.

†To whom correspondence should be addressed. E-mail: t-ohalloran@northwestern.edu

An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations

Vincent Segura^{1,2,4}, Bjarni J Vilhjálmsson^{1,3,4}, Alexander Platt^{1,3}, Arthur Korte¹, Ümit Seren¹, Quan Long¹ & Magnus Nordborg^{1,3}

Population structure causes genome-wide linkage disequilibrium between unlinked loci, leading to statistical confounding in genome-wide association studies. Mixed models have been shown to handle the confounding effects of a diffuse background of large numbers of loci of small effect well, but they do not always account for loci of larger effect. Here we propose a multi-locus mixed model as a general method for mapping complex traits in structured populations. Simulations suggest that our method outperforms existing methods in terms of power as well as false discovery rate. We apply our method to human and *Arabidopsis thaliana* data, identifying new associations and evidence for allelic heterogeneity. We also show how a priori knowledge from an *A. thaliana* linkage mapping study can be integrated into our method using a Bayesian approach. Our implementation is computationally efficient, making the analysis of large data sets ($n > 10,000$) practicable.

With the increasing availability of genomic polymorphism data, genome-wide association studies (GWAS) are becoming the default method for investigating the genetics of quantitative traits. Typically, GWAS are carried out using single-locus tests to identify associations between polymorphisms and traits in either case-control populations or cohorts. However, both study designs are subject to confounding by population structure, leading to an inflation of test statistics and a high false positive rate^{1,2}. Several methods have been proposed to address this issue, including genomic control³, structured association⁴, principal-components analysis⁵ and mixed linear models⁶. Genomic control scales the test statistics uniformly, so that the observed median test statistic equals the expected one. Even though this approach reduces the inflation of test statistics globally, it does not change the rank of the polymorphisms, as they are subject to the same correction. In the structured association and principal-component analysis approaches, population structure is taken into account by including covariates in the association model that represent

the cluster memberships and principal-component loadings of the individuals, respectively. Whereas these approaches are expected to perform well when the population structure is simple, they may perform poorly when the structure is more complex: for example, when individuals show a continuum of relatedness⁷. An additional improvement has been made with the use of mixed linear models, which are based on the insight that confounding can be caused by the genetic background of causal variants in the presence of population structure. The mixed model controls for the genetic background through a random polygenic term with a covariance structure described by a relationship matrix, so that correlations in phenotype mirror relatedness⁸, as predicted by Fisher's classical model⁹. This approach has been shown to perform well in plants, animals and humans^{6,10–12}, and methods have been developed to allow the analysis of large GWAS data sets in a reasonable amount of time^{11,13,14}.

All these approaches are based on single-locus tests combined with some kind of diffuse genomic background. However, for complex traits controlled by several large-effect loci, these approaches may not be appropriate, especially in the presence of population structure¹² (indeed, a substantial inflation of single-locus test statistics is expected for complex traits, even in the absence of population structure)¹⁵. Explicit use of multiple cofactors in the statistical model is an obvious alternative and is indeed standard in traditional linkage mapping, where both multiple-quantitative trait locus (QTL) mapping and composite interval mapping have been shown to outperform simple interval mapping^{16,17}. In GWAS, the case for including multiple loci is arguably even stronger, as the confounding effects of background loci may be present across the genome (due to linkage disequilibrium) rather than only locally (due to linkage)¹⁸. Thus, whereas conditioning on known causative factors in GWAS has typically been conducted on a local scale to help identify multiple alleles and clarify complex associations^{12,19,20}, we believe that it should be done on a genome-wide basis. As shown, conditional analysis on a genome-wide scale may well lead to higher power and a lower false discovery rate (FDR) than single-locus approaches (Fig. 1). Similarly, in the context of human genetics,

¹Gregor Mendel Institute (GMI), Austrian Academy of Sciences, Vienna, Austria. ²Institut National de la Recherche Agronomique (INRA), UR0588, Orléans, France. ³Department of Molecular and Computational Biology, University of Southern California, Los Angeles, California, USA. ⁴These authors contributed equally to this work. Correspondence should be addressed to M.N. (magnus.nordborg@gmi.oew.ac.at).

Received 16 November 2011; accepted 4 May 2012; published online 17 June 2012; doi:10.1038/ng.2314

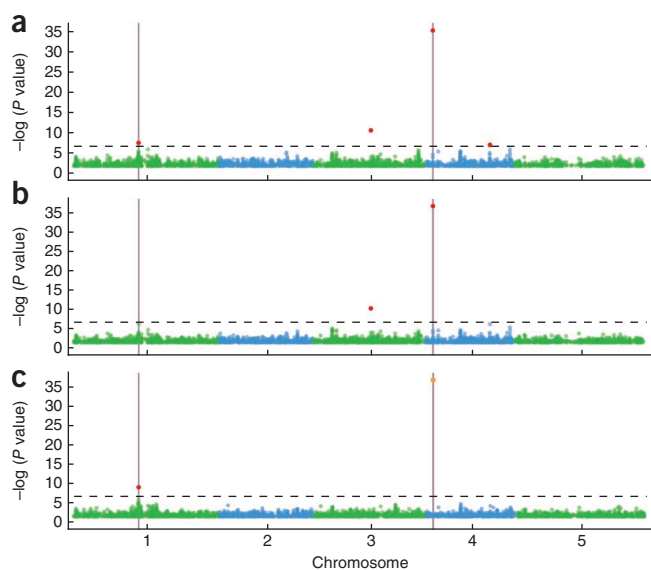


Figure 1 A GWAS for a simulated trait with two causal SNPs randomly chosen from a real *A. thaliana* SNP data set. Random error was added to the trait to fix the heritability at 25%. Causal SNPs are marked by vertical lines. **(a)** A single-SNP linear regression scan detects four significantly associated SNPs (red circles) at a Bonferroni-corrected threshold of 0.05 (dashed horizontal line). Half of these SNPs are false positives, and the other half are true positives, leading to FDR of 50% and power of 100%. **(b)** A single-SNP mixed-model^{11,14} scan eliminates one false positive but also one true positive, leading to similar FDR (50%) and decreased power of 50% compared to the model in **a**. **(c)** Adding the most significant SNP as a cofactor to the mixed model (orange circle) recovers the second causal SNP, while eliminating the last false positive, leading to the perfect scenario with FDR of 0% and power of 100%.

it has been suggested that conditioning on major-effects loci, like the major histocompatibility (MHC) region, may improve power¹¹.

However, automatically including cofactors is challenging when the number of predictors is large compared to the number of observations. This is particularly problematic in GWAS, where the number of polymorphisms (p) can reach millions but where the number of phenotyped and genotyped individuals (n) is rarely more than tens of thousands. Such ‘large p , small n ’ problems are very challenging: the model space is usually too large to explore exhaustively, and the maximum number of polymorphisms that can be fitted at a time must be less than the number of individuals. In addition, identifying causative polymorphisms by fitting more than one polymorphism at a time is complicated by the presence of linkage disequilibrium. Several approaches have been proposed to address these issues, including stepwise regression²¹ and penalized regression with different penalty functions, such as ridge regression, normal exponential gamma, elastic net and LASSO^{22–26}. These approaches have been shown to perform better than single-locus approaches, but most are either computationally unfeasible in GWAS²⁷ or do not explicitly address the problems posed by population structure. As an alternative, we propose using a simple, stepwise mixed-model regression with forward inclusion and backward elimination, which, despite being limited in terms of exploring the model space, has the advantage of being computationally efficient and therefore applicable to GWAS. To effectively address the population structure issue, we make use of an approximate version of the mixed model^{11,14} in which we re-estimate genetic and error variances at each step of the regression (Online Methods). As the variance attributed to the random polygenic term decreases when cofactors are added to the model, we propose to use the heritable variance estimate as a criterion to stop forward inclusion. Then, backward elimination is performed from the last forward model for a more thorough exploration of the model space. We evaluate various model selection criteria through simulations, which suggest that the proposed multi-locus mixed-model (MLMM) method performs well in terms of FDR and power. Finally, we show the usefulness of our approach by applying it to human and *A. thaliana* data.

RESULTS

Simulations

GWAS data were simulated by adding phenotypic effects to real genotypic data from *A. thaliana*²⁸ under two different scenarios: a 2-locus

model and a 100-locus model. For the latter, additivity was assumed, whereas, for the former, different types of interactions were explored (Online Methods).

We compared our proposed MLMM method with three other mapping methods: a single-locus approximate mixed model that corrects for population structure but does not take into account other major loci (MM)^{11,14}; a stepwise linear model that takes other major loci into account but does not correct for population structure (SWLM); and a single-locus linear model that does neither (LM). The four methods were compared in terms of their statistical power and FDR. For single-locus methods, SNPs were considered to be detected if their P values were below a defined threshold, whereas, for the multi-locus methods, detected SNPs were those belonging to the most complex model in which the marginal P values of cofactors were all below a defined threshold.

The results for the 100-locus model are shown (Fig. 2 and Supplementary Figs. 1–4), and can be summarized as follows. First, methods that use a kinship term to correct for population structure

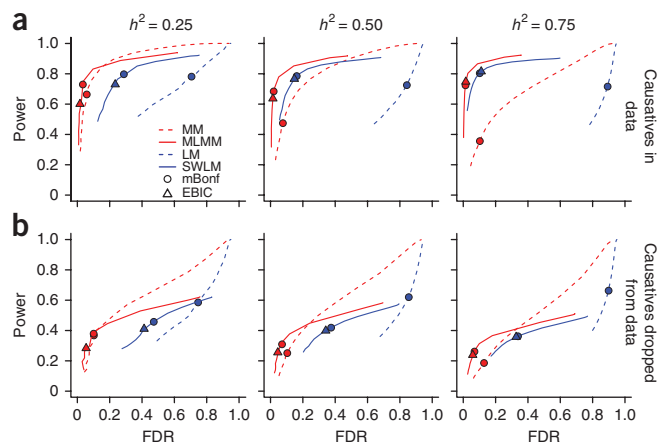


Figure 2 Power and FDR in 100-locus model simulations for four different mapping methods: LM, SWLM, MM and MLMM. **(a,b)** For the purpose of computing power and FDR, a causal SNP was considered to be detected if a SNP within 25 kb on either side was determined to have a significant association (results for other window sizes are given in Supplementary Fig. 3), and only causal SNPs that were detectable in principle (that were marginally significant at a Bonferroni-corrected threshold of 0.05 in a simple linear model) were considered. For clarity, only the backward path of the multi-locus methods (SWLM and MLMM) is shown (comparison between forward and backward paths is given in Supplementary Fig. 4). Circles and triangles represent the best-fitting model according to the mBonf and EBIC model selection criteria, respectively. Power and FDR were estimated with **(a)** and without **(b)** the causal loci included. Three phenotypic heritabilities were used in the simulations: 0.25 (left), 0.5 (middle) and 0.75 (right).

always outperform comparable methods that do not (MM and MLMM versus LM and SWLM, respectively). There is simply too much structure in these data for it to be ignored without paying a very heavy price in terms of increased FDR (Supplementary Fig. 1). Second, multi-locus methods generally outperform comparable single-locus methods (SWLM and MLMM versus LM and MM, respectively), as long as the causative sites are included in the data (Fig. 2a). The advantage increases with increasing heritability, because, under our simulation scheme, increased heritability implies more loci of large effect and, hence, greater confounding (Supplementary Figs. 1 and 2). If the causative sites themselves are excluded from the data, the single-locus mixed model (MM) may have greater power than the multi-locus version (MLMM) but only at the cost of greatly increased FDR (Fig. 2b).

The two-locus simulations allowed us to examine the advantages of including cofactors in the mixed model under several scenarios of population structure and/or epistasis (Online Methods). Regardless of the scenario considered, MLMM consistently performed at least as well as the other methods when restricted to a small FDR (Fig. 1 and Supplementary Fig. 5). When two causal sites were chosen at random, the improvement in power observed for MLMM over that in the single-marker MM was almost entirely attributed to increased power to detect the second causal site (Supplementary Fig. 6).

A serious problem when employing multi-locus models is knowing how many loci to include. We propose two model selection criteria: the extended Bayesian information criterion (EBIC)²⁹ and the

multiple-Bonferroni criterion (mBonf), defined as the largest model in which all cofactors have a *P* value below a Bonferroni-corrected threshold (we used a threshold of 0.05). Our simulations showed that both criteria are consistent in bounding the FDR for the MLMM method, regardless of the simulation scenario, with EBIC being slightly more stringent than mBonf (Fig. 2 and Supplementary Fig. 5). In addition, the genome-wide *P* values in the models selected by both criteria were uniformly distributed, showing the ability of mixed models to control confounding by population structure in a multi-locus setting (Supplementary Fig. 1). Furthermore, both criteria performed appropriately in extreme scenarios where there was no detectable signal in the data, as might occur when an external confounding variable interacts nonlinearly with a single causal locus¹⁸. In this case, MLMM with one of the proposed criteria correctly selected a model without any SNPs, whereas the other methods tested would identify only false positives (Supplementary Fig. 5). In summary, MLMM with the conservative FDR provided by the proposed model selection criteria consistently outperformed the other methods in all scenarios that we examined.

For completeness, we also compared MLMM to other single-locus mixed-model implementations, including the exact mixed model³⁰ and the approximate mixed model with compression¹⁴, as they have been shown to perform better than the approximate method. These methods did indeed perform slightly better than the approximate method in our simulations but were still far from the performance achieved by MLMM (Supplementary Fig. 7).

Table 1 SNPs identified in multi-locus mixed-model analysis of NFBC1966 traits

SNP	Chr.	Position	Gene	<i>P</i> value		Previous identification	
				EBIC	mBonf	Sabatti <i>et al.</i>	Kang <i>et al.</i>
Associated with triglyceride levels (mM)							
rs673548	2	21091049	<i>APOB</i>		5.1×10^{-8}	+	+
rs1260326	2	27584444	<i>GCKR</i>	1.5×10^{-10}	7.9×10^{-11}	+	+
rs10096633	8	19875201	<i>LPL</i>	1.6×10^{-8}	2.4×10^{-8}	+	+
Associated with HDL levels (mM)							
rs1532085	15	56470658	<i>LIPC</i>	9.2×10^{-12}	8.0×10^{-12}	+	+
rs3764261	16	55550825	<i>CETP</i>	2.7×10^{-32}	3.7×10^{-23}	+	+
rs7499892	16	55564091	<i>CETP</i>		9.5×10^{-8}	-	-
rs255049	16	66570972	<i>LCAT</i>	1.3×10^{-8}	4.8×10^{-8}	+	+
rs1800961	20	42475778	<i>HNF4A</i>		1.5×10^{-7}	-	-
Associated with LDL levels (mM)							
rs646776	1	109620053	<i>CELSR2</i>	4.2×10^{-16}	4.2×10^{-16}	+	+
rs693	2	21085700	<i>APOB</i>	7.1×10^{-12}	7.1×10^{-12}	+	+
rs11668477	19	11056030	<i>LDLR</i>	1.0×10^{-9}	1.0×10^{-9}	+	+
rs157580	19	50087106	<i>TOMM40-APOE</i>	2.2×10^{-17}	2.2×10^{-17}	+	-
rs405509	19	50100676	<i>TOMM40-APOE</i>	1.3×10^{-12}	1.3×10^{-12}	-	-
Associated with CRP levels (mM)							
rs2369146	1	157934819	<i>CRP</i>	4.5×10^{-9}	2.8×10^{-9}	-	-
rs2794520	1	157945440	<i>CRP</i>	1.1×10^{-29}	6.6×10^{-30}	+	+
rs2650000	12	119873345	<i>HNF1A</i>	1.3×10^{-12}	1.0×10^{-12}	+	+
rs8106922	19	50093506	<i>TOMM40-APOE</i>		1.6×10^{-12}	-	-
rs439401	19	50106291	<i>TOMM40-APOE</i>		2.2×10^{-9}	-	-
Associated with glucose levels (mM)							
rs560887	2	169471394	<i>G6PC2</i>	2.2×10^{-13}	2.2×10^{-13}	+	+
rs2971671	7	44177862	<i>GCK</i>	3.2×10^{-9}	3.2×10^{-9}	-	+
rs3847554	11	92308474	<i>MTNR1B</i>	4.7×10^{-11}	4.7×10^{-11}	- ^a	-
Associated with SBP							
rs782602	2	55702813	<i>SMEK2</i>		1.4×10^{-7}	-	- ^b

Models were selected using either EBIC or mBonf. Chr., chromosome.

^aThis SNP was not reported by Sabatti *et al.*, but they reported two other SNPs located in the same gene. ^bKang *et al.* did not report this association because they used a *P*-value threshold slightly more stringent than the Bonferroni-corrected threshold of 0.05 used for mBonf.



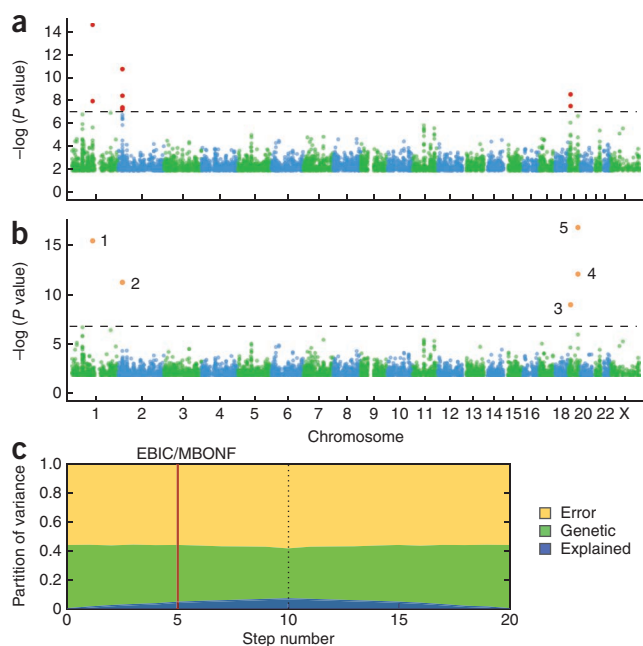


Figure 3 GWAS for LDL levels in the NFBC1966 data set. (a) A single-locus mixed model identifies seven SNPs in three genes (red circles) at a Bonferroni-corrected threshold of 0.05 (dashed horizontal line). (b) MLMM identifies five SNPs in four genes (orange circles numbered in the order in which they were included in the model). (c) Partition of variance at each step of MLMM (ten forward and ten backward) into variance explained by the SNPs included in the model, kinship and noise.

Application to a human data set

To show the feasibility as well as the usefulness of MLMM, we applied it to a previously published data set of metabolic traits in the Northern Finland Birth Cohort (NFBC1966)³¹. The data were previously reanalyzed to show the usefulness of the mixed model¹¹, and we used the same settings for mixed-model estimation here. The SNPs identified using MLMM are listed in **Table 1**. As predicted by our simulations, EBIC was more stringent than mBonf, resulting in the selection of models that were either similar to or nested within the models selected by mBonf. Using the less-conservative mBonf criterion, we identified all the associations previously detected with the single-locus mixed model¹¹ and nine additional associations. Of the newly identified association signals, three were located near genes previously reported using the same data³¹ (two in the *TOMM40-APOE* cluster for low-density lipoprotein (LDL) levels and one in *MTNR1B* for glucose levels), and four were located in gene regions not previously reported with this data set (one in *HNF4A* for high-density lipoprotein (HDL) levels, one in *SMEK2* for systolic blood pressure (SBP) and two in the *TOMM40-APOE* cluster for C-reactive protein (CRP) levels). The remaining two associations were additional SNPs in genes that had already been reported (*CETP* for HDL and *CRP* for CRP levels). The detected association in *HNF4A* for HDL levels (rs1800961) in a gene

Figure 4 GWAS for sodium accumulation in *A. thaliana*. (a) A single-locus mixed model identifies a strong peak of significantly associated SNPs on chromosome 4 (red circles) at a Bonferroni-corrected threshold of 0.05 (dashed horizontal line). (b) MLMM identifies three SNPs (orange circles numbered in the order in which they were included in the model). (c) Partition of variance at each step of MLMM (eight forward and eight backward) into variance explained by the SNPs included in the model, kinship and noise.

region not previously reported with this data set has been replicated in two meta-analyses of 30,714 and 99,900 individuals each^{32,33}.

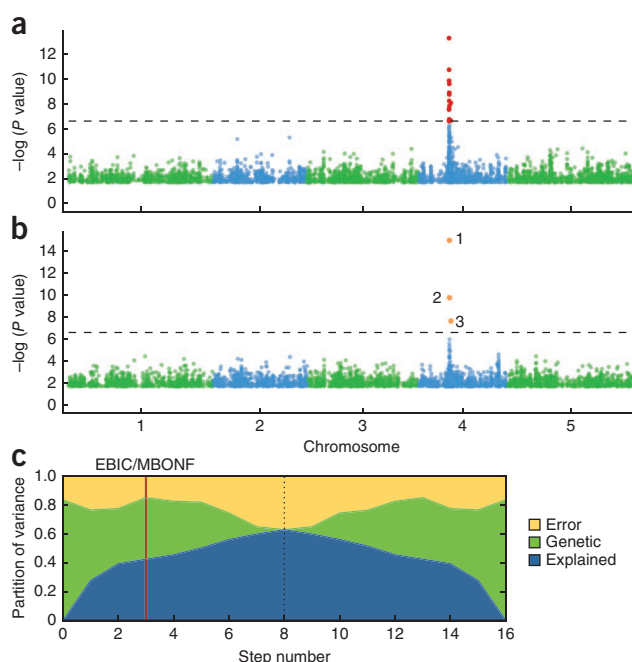
Multiple SNPs with significant association signals within or near a single gene suggest either allelic heterogeneity or the presence of an untyped causal variant that is partially represented by multiple SNPs (or both). In the case of the associations located in the *TOMM40-APOE* cluster (for both LDL and CRP levels), we observed a marked decrease in the *P* values for the two selected SNPs when they were both included in the model (**Fig. 3** and **Supplementary Fig. 8**), which presumably explains why they were not identified using the single-locus mixed model. This type of situation is expected when loci mask each other, for example, when alleles of compensatory effect are correlated, as seems to be the case here ($R^2 = 0.33$ and 0.25 for LDL and CRP levels, respectively).

We show the percentage of variance explained by the SNPs included in the model and the percentages of unexplained genetic and residual variance at the different steps of the MLMM for LDL levels (**Fig. 3** and **Supplementary Fig. 9**). It is notable that most of the heritable phenotypic variation remains unexplained.

Application to an *A. thaliana* data set

Sodium accumulation in the leaves of *A. thaliana* has been shown to be strongly associated with genotype and expression levels of the Na^+ transporter *AtHKT1;1* (ref. 34). In particular, a SNP located in the first exon of the gene (chromosome 4: 6,392,280) shows a highly significant association (P value = 6.33×10^{-14} using an approximate mixed model). We reanalyzed these data using MLMM and found that the sole SNP previously reported³⁴ only explains part of the signal in the associated region (**Fig. 4**).

Instead, the optimal model obtained with MLMM (according to both EBIC and mBonf) included three SNPs, which together explained 42.3% of the phenotypic variation. This model included the previously reported SNP, which explained 27.7% of the variation, and a second SNP only 22 kb away from the gene, suggesting that there might be multiple causal variants in the gene. To further investigate the associations in this particular region, we applied our method locally, using only the 508 SNPs located within 100 kb of the gene. Using EBIC,



six SNPs were included in the model, all within 25 kb of *AtHKT1;1*, which explained 52.6% of the phenotypic variation (**Supplementary Fig. 10**), leaving 20.5% of the heritable fraction of the total variance unexplained. As noted, this suggests either allelic heterogeneity or the presence of one or more untyped causal variants. However, as the largest possible fraction of variance explained by a single binary SNP (which would have a minor allele frequency of 0.32) is 47.6%, we conclude that there is evidence for allelic heterogeneity in this case.

DISCUSSION

The problem of population structure in GWAS is best viewed as one of model mis-specification. Single-locus tests of association are the wrong model to use in cases where the trait is not attributable to a single locus. Ignoring the genetic background may be defensible in some circumstances but is clearly not when causative alleles are correlated across loci due to population structure and/or selection¹², resulting in biased estimates of effect sizes. The problem has long been recognized by animal breeders, who developed a mixed linear model to reduce bias⁸. This approach works well but assumes that the phenotypic covariance between individuals can be predicted by their relatedness, as estimated by genotypes at SNPs across the genome. As demonstrated by Fisher close to 100 years ago⁹, this approximation is reasonable if the genetic background is sufficiently smooth, but it is clear that loci of relatively large effect may make this approach invalid¹⁸. We therefore propose to extend the mixed model for GWAS to include multiple loci, in parallel to what is routinely done in QTL linkage mapping^{16,17}.

Our proposed method includes significant effects in the model via a forward-backward stepwise approach, while re-estimating the variance components of the model at each step. If the fixed effects included are real, they can reduce the unexplained heritable variance and effectively lower the restraints posed by the mixed model on other markers that correlate with population structure. As demonstrated by simulations, our MLM model implementation shows promising performance in terms of power and FDR in comparison with a single-marker scan and a stepwise linear regression, especially when applying a conservative threshold, which can be achieved with one of the proposed model quality criteria. In particular, MLM model performed much better than the other methods tested for structured samples and traits involving several loci with moderate to large effect.

Applying MLM model to real data from humans and *A. thaliana*, we identified interesting new associations as well as evidence for allelic heterogeneity. Indeed, as it includes multiple loci in the model, MLM model helps identify evidence for allelic heterogeneity in addition to interactions, although it is difficult to exclude the possibility that multiple associated SNPs within a region are detected because of partial linkage disequilibrium with an untyped causal variant^{12,18,20}. However, with the rapid development of DNA sequencing³⁵, it is increasingly likely that causal variants will be typed. As seen in our simulations, all tested methods, especially MLM model, will benefit greatly from this. Applied here to the analysis of quantitative traits, MLM model can also be applied to the study of disease heritability. Indeed, it is possible to analyze a disease phenotype with an approximate mixed model by considering a binary quantitative response corresponding to case-control status¹¹. MLM model partitions the phenotypic variance into genetic, random and explained variance at each step, suggesting a natural stopping criterion (genetic variance of 0) for including cofactors. This allows the user to obtain estimates of the explained and unexplained heritable variance, as well as gives insights into trait architecture.

MLM model is far from a panacea, however. The greedy forward-backward inclusion of SNPs is clearly limited in exploring the huge model space.

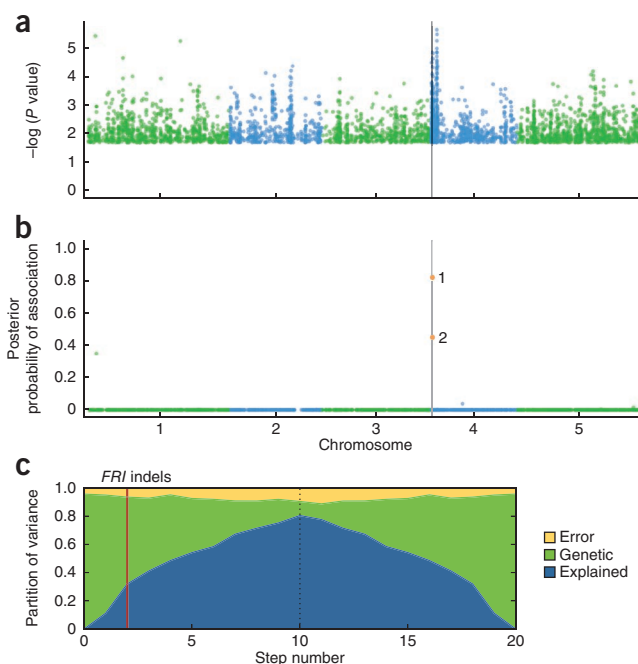


Figure 5 An example of Bayesian MLM model for the analysis of *FLC* expression in *A. thaliana*. **(a)** An approximate mixed-model scan for *FLC* expression, with the *FRIGIDA* gene marked by a vertical line. **(b)** The posterior probability of association scan after the Bayesian MLM model has included two loci in the model (orange circles), which incidentally are the two causative indels previously identified. **(c)** Partition of phenotypic variance for each forward inclusion (ten steps) and backward elimination (ten steps after the dashed line). The vertical red line marks the model with the two causative indels included in the model.

More sophisticated algorithms, like LASSO³⁶, are worth exploring. However, similar to other penalized methods, LASSO typically assumes independence between markers, which would not be appropriate for structured data. In the context of structured data, LASSO might give a large effect size to a marker that is in linkage disequilibrium with many other markers, whereas a mixed model would down-weight such markers. A potential improvement on this would be to use LASSO in conjunction with a mixed model²⁶. Although this approach is potentially very promising, it is currently too computationally demanding for GWAS data sets. Another promising approach is resample model averaging³⁷, which has been applied successfully to joint linkage association analysis³⁸. However, it is important to realize that the problem is fundamentally very difficult. For example, we have previously shown that linkage disequilibrium between two known causal alleles of the *A. thaliana* flowering locus *FRIGIDA* (*FRI*) and the genomic background give rise to a very complicated pattern of association in a GWAS of *FLOWERING LOCUS C* (*FLC*) expression¹². None of the methods tested here identified the causal sites. This is not unexpected, as there are many spurious one- and two-locus models that fit the data better than those involving the true causal loci. In cases like this, we think it is unlikely that progress will be made without independent data to help prioritize variants. As MLM model is based on a linear model, it can easily be extended for Bayesian analysis^{39,40} and allows for the integration of previous knowledge into the model. Indeed, returning to the *FLC* example, by placing a 100-fold prior on all markers within 10 kb of *FRI*, we allow MLM model to include the two known causal variations as the first two cofactors in the model, showing how priors can help identify causal loci and improving the model (**Fig. 5**).

URLs. INRA MIGALE platform, <http://migale.jouy.inra.fr/>; R version of MLMM, <https://cynin.gmi.oew.ac.at/home/resources/mlmm>; Python version of MLMM, <https://github.com/bvilhjal/mixmogam>; Scientific Tools for Python (SciPy) package, <http://www.scipy.org/>.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Supplementary information is available in the online version of the paper.

ACKNOWLEDGMENTS

We acknowledge the NFBC1966 Study investigators for allowing us to use their phenotype and genotype data in our study. The NFBC1966 Study is conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with the Broad Institute, the University of California, Los Angeles (UCLA), the University of Oulu and the National Institute for Health and Welfare in Finland. This manuscript was not prepared in collaboration with the investigators from the NFBC1966 Study and does not necessarily reflect the opinions or views of these investigators or those at the collaborating institutes. We thank N.B. Freimer and S.K. Service for their help in pre-processing the NFBC1966 data. We would also like to thank P. Forai for excellent information technology and cluster support at GMI, the INRA MIGALE bioinformatics platform for additional computational resources and D.V. Conti, D.J. Balding and S. Srivastava for useful discussions on the topic. Finally, we would like to thank the anonymous reviewers for their helpful comments on the manuscript. This work was supported by grants from the Ecologie des Forêts, Prairies et milieux Aquatiques (EPPA) department of INRA to V.S. and Deutsche Forschungsgemeinschaft (DFG) to A.K. and by grants from the US National Institutes of Health (P50 HG002790) and the European Union Framework Programme 7 (TransPLANT, grant agreement 283496) to M.N., as well as by the Austrian Academy of Sciences through GMI.

AUTHOR CONTRIBUTIONS

All authors contributed to designing the study. V.S. and B.J.V. ran the simulations and analyzed the data. V.S., B.J.V. and M.N. wrote the manuscript with input from A.P., A.K., Ü.S. and Q.L.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at <http://www.nature.com/doi/10.1038/ng.2314>.
Reprints and permissions information is available online at <http://www.nature.com/reprints/index.html>.

1. Cardon, L.R. & Palmer, L.J. Population stratification and spurious allelic association. *Lancet* **361**, 598–604 (2003).
2. Marchini, J., Cardon, L.R., Phillips, M.S. & Donnelly, P. The effects of human population structure on large genetic association studies. *Nat. Genet.* **36**, 512–517 (2004).
3. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55**, 997–1004 (1999).
4. Pritchard, J.K., Stephens, M., Rosenberg, N.A. & Donnelly, P. Association mapping in structured populations. *Am. J. Hum. Genet.* **67**, 170–181 (2000).
5. Price, A.L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
6. Yu, J. *et al.* A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.* **38**, 203–208 (2006).
7. Zhao, K. *et al.* An *Arabidopsis* example of association mapping in structured samples. *PLoS Genet.* **3**, e4 (2007).
8. Henderson, C.R. *Application of Linear Models in Animal Breeding* (University of Guelph, Guelph, Canada, 1984).
9. Fisher, R.A. The correlation between relatives on the supposition of Mendelian inheritance. *Trans. R. Soc. Edinb.* **52**, 399–433 (1918).
10. Kang, H.M. *et al.* Efficient control of population structure in model organism association mapping. *Genetics* **178**, 1709–1723 (2008).

11. Kang, H.M. *et al.* Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* **42**, 348–354 (2010).
12. Atwell, S. *et al.* Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. *Nature* **465**, 627–631 (2010).
13. Aulchenko, Y.S., de Koning, D.J. & Haley, C. Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. *Genetics* **177**, 577–585 (2007).
14. Zhang, Z. *et al.* Mixed linear model approach adapted for genome-wide association studies. *Nat. Genet.* **42**, 355–360 (2010).
15. Yang, J. *et al.* Genomic inflation factors under polygenic inheritance. *Eur. J. Hum. Genet.* **19**, 807–812 (2011).
16. Jansen, R.C. Interval mapping of multiple quantitative trait loci. *Genetics* **135**, 205–211 (1993).
17. Zeng, Z.B. Precision mapping of quantitative trait loci. *Genetics* **136**, 1457–1468 (1994).
18. Platt, A., Vilhjalmsón, B.J. & Nordborg, M. Conditions under which genome-wide association studies will be positively misleading. *Genetics* **186**, 1045–1052 (2010).
19. Allen, A.S., Satten, G.A., Bray, S.L., Dudbridge, F. & Epstein, M.P. Fast and robust association tests for untyped SNPs in case-control studies. *Hum. Hered.* **70**, 167–176 (2010).
20. Dickson, S.P., Wang, K., Krantz, I., Hakonarson, H. & Goldstein, D.B. Rare variants create synthetic genome-wide associations. *PLoS Biol.* **8**, e1000294 (2010).
21. Cordell, H.J. & Clayton, D.G. A unified stepwise regression procedure for evaluating the relative effects of polymorphisms within a gene using case/control or family data: application to HLA in type 1 diabetes. *Am. J. Hum. Genet.* **70**, 124–141 (2002).
22. Hoggart, C.J., Whittaker, J.C., De Iorio, M. & Balding, D.J. Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. *PLoS Genet.* **4**, e1000130 (2008).
23. Malo, N., Libiger, O. & Schork, N.J. Accommodating linkage disequilibrium in genetic-association analyses via ridge regression. *Am. J. Hum. Genet.* **82**, 375–385 (2008).
24. Croiseau, P. & Cordell, H.J. Analysis of North American Rheumatoid Arthritis Consortium data using a penalized logistic regression approach. *BMC Proc.* **3**, S61 (2009).
25. Cho, S. *et al.* Joint identification of multiple genetic variants via elastic-net variable selection in a genome-wide association analysis. *Ann. Hum. Genet.* **74**, 416–428 (2010).
26. Wang, D., Eskridge, K.M. & Crossa, J. Identifying QTLs and epistasis in structured plant populations using adaptive mixed LASSO. *J. Agric. Biol. Environ. Stat.* **16**, 170–184 (2011).
27. Ayers, K.L. & Cordell, H.J. SNP selection in genome-wide and candidate gene studies via penalized logistic regression. *Genet. Epidemiol.* **34**, 879–891 (2010).
28. Horton, M.W. *et al.* Genome-wide patterns of genetic variation in worldwide *Arabidopsis thaliana* accessions from the RegMap panel. *Nat. Genet.* **44**, 212–216 (2012).
29. Chen, J.H. & Chen, Z.H. Extended Bayesian information criteria for model selection with large model spaces. *Biometrika* **95**, 759–771 (2008).
30. Astle, W. & Balding, D.J. Population structure and cryptic relatedness in genetic association studies. *Stat. Sci.* **24**, 451–471 (2009).
31. Sabatti, C. *et al.* Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nat. Genet.* **41**, 35–46 (2009).
32. Kathiresan, S. *et al.* Common variants at 30 loci contribute to polygenic dyslipidemia. *Nat. Genet.* **41**, 56–65 (2009).
33. Teslovich, T.M. *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**, 707–713 (2010).
34. Baxter, I. *et al.* A coastal cline in sodium accumulation in *Arabidopsis thaliana* is driven by natural variation of the sodium transporter *AtHKT1;1*. *PLoS Genet.* **6**, e1001193 (2010).
35. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
36. Tibshirani, R. Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc., B* **58**, 267–288 (1996).
37. Valdar, W., Holmes, C.C., Mott, R. & Flint, J. Mapping in structured populations by resample model averaging. *Genetics* **182**, 1263–1277 (2009).
38. Tian, F. *et al.* Genome-wide association study of leaf architecture in the maize nested association mapping population. *Nat. Genet.* **43**, 159–162 (2011).
39. Stephens, M. & Balding, D.J. Bayesian statistical methods for genetic association studies. *Nat. Rev. Genet.* **10**, 681–690 (2009).
40. Servin, B. & Stephens, M. Imputation-based analysis of association studies: candidate regions and quantitative traits. *PLoS Genet.* **3**, e114 (2007).



ONLINE METHODS

Data. Both *A. thaliana* and human data were used for the examples. The genotype data for *A. thaliana* included 1,307 individual plants genotyped at 214,051 SNPs using a 250K Affymetrix SNP chip²⁸. The two *A. thaliana* phenotype data sets used were (i) sodium levels averaged over 6 replicates of 342 accessions³⁴ and (ii) *FLOWERING LOCUS C (FLC)* expression measured in 166 accessions¹². For *FLC* expression, the genotype data used were the same as those previously described¹², which come from a subset of the 1,307 individual plants and contain 216,130 markers, including 3 indels within or near the *FRIGIDA (FRI)* gene. For priors, we gave every marker that was within 10 kb of the *FRI* gene a 100-fold greater prior than the base prior. We then scaled them so that the sum of the priors over all the SNPs was 1.

The human data set used was the 1966 North Finland Birth Cohort NFBC1966 composed of 5,402 individuals having both phenotypic and genotypic data³¹. Phenotypic data consisted of measures for 10 quantitative traits, and genotypic data were available for 368,177 SNP markers. We were able to obtain the exact same data set, including 5,326 individuals and 331,475 SNPs after filtering, that was used previously¹¹. The proportion of missing genotypes was <1%; we imputed the missing genotypes with the corresponding average per SNP to speed up the mixed-model computations.

Simulations. Using the *A. thaliana* genotypic data²⁸, we simulated two types of traits: simple ones controlled by one or two causal loci and complex ones controlled by 100 loci. For the simple traits, two randomly chosen SNPs or one randomly chosen SNP and one binary latent variable were used to generate phenotypes with three phenotypic models (additive, and/or, xor; **Supplementary Table 1**). The latent binary variable was designed by dividing the accessions in half on the basis of their latitude of origin, which we refer to as the latent north-south variable, to generate substantial covariance between the phenotypes and population structure. An additional random deviation was added, drawn from a multivariate normal distribution having a mean of zero and a scaled identity matrix as covariance to fix the trait heritability to 0.1. We simulated 1,000 phenotypes for each simulation type (two causative SNPs or one causative SNP and the latent binary variable), phenotypic model and phenotypic heritability. For complex traits, we used an additive model with 100 randomly sampled SNPs having effect sizes drawn from an exponential distribution with a rate of 1. An additional random deviation was added, drawn from a normal distribution with a mean of zero and scaled identity matrix as covariance matrix to fix the trait heritability to 0.25, 0.5 and 0.75. For each phenotypic heritability, 500 phenotypes were simulated. All simulated phenotypes have been analyzed with the four methods presented in the main text. For completeness, another single-locus approximate mixed model was used to analyze the phenotypes simulated under the 100-locus model. To control some potential confounding from population structure that was not accounted for by the random term, this approach uses as covariates the ten first principal components from a principal-component analysis of the standardized genotypic data. As no obvious difference was observed between this additional approach and the approximate mixed model, only the latter was presented (**Supplementary Fig. 11**).

Linear mixed model. Following Fisher's⁹ polygenic model and adopting similar notation as was used previously⁴¹, the phenotypic value of the *i*th individual can be denoted as

$$y_i = \mu + \sum_{j=1}^m x_{ij}a_j + e_i$$

where *m* is the total number of causal loci, x_{ij} is the genotype (coded in numerical terms) of the *j*th causal locus to the *i*th individual, a_j is the effect size of the *j*th locus and e_i is the error. If we assume that there are a large number of independent causal loci and that their effects are drawn from a Gaussian distribution (Fisher's infinitesimal model), we can sum them and approximate them with a Gaussian random variable. We therefore modeled the trait using a mixed model⁸, where the phenotype can be denoted in vector notation as $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{g} + \mathbf{e}$, where \mathbf{X} is a matrix of fixed effects (for example, SNPs), $\boldsymbol{\beta}$ is a vector of effect sizes, \mathbf{g} is a vector of random polygenic effects with covariance matrix $\sigma_g^2 K^*$ and \mathbf{e} is a vector of random independent effects with variance σ_e^2 modelling the residual error. Both the random terms, \mathbf{g} and \mathbf{e} , are assumed to have a

Gaussian distribution with mean of 0. Here, K^* denotes the adjusted kinship matrix, where the loci included as fixed effects are excluded from kinship matrix estimation. If $M \gg n$, where M is the number of causal loci and n is the number of individuals, then $K^* \approx K$. Different assumptions lead to different kinship matrices that can be used for the mixed model as described in the **Supplementary Note**.

Multiple loci mixed model. We used forward-backward stepwise linear mixed-model regression, where the variance components σ_g^2 and σ_e^2 are estimated before each step. The variance estimates are used to obtain generalized least-square (GLS) effect size estimates and F-test *P* values for each SNP. The SNP with the most significant association is then added to the model as a cofactor for the next step, and the *P* values for all cofactors are re-estimated together with the variance components. For stopping criteria for the forward regression, we suggest stopping when the $\sigma_g^2 / \text{var}(\mathbf{y})$ estimate is close to zero or when a maximum number of forward steps is reached. After stopping the forward stepwise regression, a backward stepwise regression is performed by dropping the least significant cofactor in the model at each step. The variance components and *P* values of all cofactors are again re-estimated at each step. For variance component estimation at each forward and backward step, the markers included as cofactors in the model can be excluded from the kinship matrix calculation, although we did not do this, as their effect on kinship is arguably negligible.

We made use of the Gram-Schmidt process⁴¹, which makes each step as fast as the first one when $M \gg n$ (when the number of SNPs is much greater than the number of individuals). At each step, we obtained the QR decomposition of the cofactor matrix to obtain the *Q* matrix, and we used this to calculate the marginal inverse-variance matrix as

$$M^{-1} = (I - Q'Q)'V^{-1}$$

where $V = \sigma_g^2 K + \sigma_e^2 I$ is the covariance matrix estimated at each step.

We explored several model selection criteria to select the most appropriate model. The classic Bayesian information criterion (BIC) is too tolerant in the context of GWAS, allowing for too many loci in the model, and is therefore not recommended. As an alternative, we used extended BIC, initially defined as the BIC penalized by the model space dimension²⁹. We also propose and define a new criterion, the multiple Bonferroni criterion (mBonf), which selects the model with the most loci that all have *P* values below the Bonferroni threshold. This criterion enables the user to specify the *P*-value threshold if one wants to allow for a higher FDR or restrict to a lower one. The computational complexity of our implementation is described in the **Supplementary Note**.

Employing priors on loci. As described³⁹, it is possible to employ priors on loci in a Bayesian model, where the Bayes factor is calculated for each locus. Calculating the Bayes factor, however, is not always easy, as it requires integrating out the model parameters that have some specified prior distributions. In our case, the model parameters of interest were the effect sizes of the loci in the model. A rough approximation can be achieved using the Schwarz criterion, which allowed us to avoid defining priors on the effect sizes and evaluating the integral⁴². We define the approximate Bayes factor (ABF) as

$$\log ABF = \log P(D | \boldsymbol{\beta}, M_1) - \log P(D | \boldsymbol{\beta}, M_0) - \frac{1}{2}(d_1 - d_0) \log n$$

where n is the number of individuals, D is the observed data, M_i is the *i*th model and d_i is the degree of freedom in the *i*th model. Using this approximation together with a prior probability π for the causal locus, we define the approximate posterior probability of association (APPA).

$$APPA = \frac{ABF \times \pi}{1 - \pi(1 + ABF)}$$

We note that this quantity should be treated more as a score than a probability, as it is a rough estimate of the actual probability.

Software availability. MLMM has been implemented in two programming languages, Python and R, which have been made available (see URLs). The R

version relies on the original EMMA implementation¹⁰. The Python version relies heavily on the SciPy package, which can be compiled with different basic linear algebra subprograms (BLAS) versions, including GotoBLAS and the Intel Math Kernel Library (MKL).

41. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning* (Springer, New York, 2009).
42. Kass, R.E. & Raftery, A.E. Bayes Factors. *J. Am. Stat. Assoc.* **90**, 773–795 (1995).

A mixed-model approach for genome-wide association studies of correlated traits in structured populations

Arthur Korte^{1,4}, Bjarni J Vilhjálmsson^{1,2,4}, Vincent Segura^{1,3,4}, Alexander Platt^{1,2}, Quan Long¹ & Magnus Nordborg^{1,2}

Genome-wide association studies (GWAS) are a standard approach for studying the genetics of natural variation. A major concern in GWAS is the need to account for the complicated dependence structure of the data, both between loci as well as between individuals. Mixed models have emerged as a general and flexible approach for correcting for population structure in GWAS. Here, we extend this linear mixed-model approach to carry out GWAS of correlated phenotypes, deriving a fully parameterized multi-trait mixed model (MTMM) that considers both the within-trait and between-trait variance components simultaneously for multiple traits. We apply this to data from a human cohort for correlated blood lipid traits from the Northern Finland Birth Cohort 1966 and show greatly increased power to detect pleiotropic loci that affect more than one blood lipid trait. We also apply this approach to an *Arabidopsis thaliana* data set for flowering measurements in two different locations, identifying loci whose effect depends on the environment.

Most GWAS to date have been conducted using the simplest possible statistical model: a single-locus test of association between a binary SNP genotype and a single phenotype. Given that most traits of interest are multifactorial, this clearly amounts to model misspecification, and the resulting danger of biased results whenever there is a lack of independent (linkage disequilibrium) between causal loci (for example, due to population structure) is well known^{1–3}. Much less attention has been devoted to the fact that phenotypes may also be correlated. Whenever multiple measurements are taken from individuals, the resulting phenotypes will be correlated because of pleiotropy, which is of direct interest, as well as shared environment and linkage disequilibrium, which are usually confounding factors. Taking these correlations into account is important, not only because of the importance of understanding pleiotropy, but also because we may expect increased power compared to marginal analyses. Intuitively, correlated traits amount to a form of replication. The importance of correlated phenotypes becomes even clearer when we consider measurements across environments. The canonical example here is an agricultural field experiment using inbred lines, a setting in which no one would consider

analyzing phenotypes from different environments independently of each other because the whole point of the study is to separate genetic from environmental effects and identify genotype-environment interactions. In human genetics, disentangling genetic and environmental effects is also of obvious interest, although much more challenging, as the environment usually cannot be experimentally manipulated⁴.

There is a long history of multi-trait models in quantitative genetics^{5–9}, but these methods have rarely been applied to GWAS. In this paper, we show how a standard linear mixed model from animal breeding¹⁰ may be used to model correlated traits, while at the same time correcting for dependence among loci (for example, due to population structure). As designs like cohort studies become more prevalent, the need for modeling correlated traits as well as population structure will grow^{2,11,12}, and the same is true for the increasing number of nonhuman GWAS^{13–17}.

The mixed model, which handles population structure by estimating the phenotypic covariance that is due to genetic relatedness—or kinship—between individuals, has previously been shown to perform well in GWAS^{2,13,18–22}. Here, we extend this approach to handle correlated phenotypes by deriving a fully parameterized multi-trait mixed model (MTMM) that considers both the within-trait and between-trait variance components simultaneously for multiple traits (Online Methods), implementing it for GWAS. The idea is not new^{23–27}, but it has never been applied for association mapping on a genome-wide scale. Alternative approaches for GWAS analysis at multiple traits exist, but they generally are unable to control for population structure^{28,29}, and often are not applicable to genome-wide data.

We validate our approach using extensive simulations based on available SNP data from *A. thaliana*³⁰, showing that our model increases power to detect associations while controlling the false discovery rate. We then demonstrate its usefulness by considering correlated blood lipid traits from the Northern Finland Birth Cohort 1966 (NFBC1966)³¹ and environmental plasticity in an *A. thaliana* data set that contains flowering measurements for two simulated growth seasons in two different locations³². Finally, we discuss the usefulness of this approach, not only in terms of increasing power to detect associations, but also in terms of understanding the basic genetic architecture of the phenotypes.

¹Gregor Mendel Institute, Austrian Academy of Sciences, Vienna, Austria. ²Department of Molecular and Computational Biology, University of Southern California, Los Angeles, California, USA. ³Institut National de la Recherche Agronomique (INRA), UR0588, Orléans, France. ⁴These authors contributed equally to this work. Correspondence should be addressed to M.N. (magnus.nordborg@gmi.oew.ac.at).

Received 17 January; accepted 5 July; published online 19 August 2012; doi:10.1038/ng.2376

RESULTS

Simulations

Pairs of correlated phenotypes were simulated by adding phenotypic effects to genome-wide SNP data from *A. thaliana*³⁰. A single randomly selected SNP was set to account for up to 2% of the phenotypic variance, but with the possibility of different effects in each of the two phenotypes. In addition, 10,000 SNPs were given much smaller effects to simulate the genetic background. A randomly chosen fraction of these background SNPs was shared between the two phenotypes, allowing for variation in the degree of phenotypic correlation (Online Methods and **Supplementary Fig. 1**).

We compared our ability to identify the focal locus using MTMM and marginal, single-trait analyses (using the smallest *P* value from the latter to ensure a fair comparison). Three different tests were used: a ‘full test’ that compares the full model, including the effect of the marker genotype and its interaction, with a model that includes neither; an ‘interaction effect test’ that compares the full model to one that does not include interaction; and a ‘common effect test’ that compares a model with a marker genotype to one without (see Online Methods for details). As expected, the results depended greatly on the effect of the focal polymorphism (**Fig. 1**). When this polymorphism had the same effect in both phenotypes (positive pleiotropy or a common effect across environments; **Fig. 1a**), MTMM performed slightly better than the single-trait mixed model, regardless of whether we tested for full model fit or just for a common effect (**Fig. 1e**). The reason for this is the increased power that results from analyzing the traits together. There is no rationale to testing for an interaction effect, as no interaction exists.

When the effect of the polymorphism is slightly weaker in one trait or environment (**Fig. 1b**), testing for a full model fit using MTMM again outperformed single-trait analyses (**Fig. 1f**). Testing only for a common or interaction effect using MTMM is also less effective. Although an interaction effect now exists, it is too weak to be detected. However, as the strength of the interaction effect increases (**Fig. 1c,d**), it becomes possible to detect directly, and the relative advantage of using MTMM increases markedly (**Fig. 1g,h**).

An alternative to carrying out two marginal single-trait analyses might be to combine the phenotypes, for example, by fitting the principal components of the traits or their sum or difference. We tested the latter, and, as might be expected, this approach worked very well when the focal SNP had exactly the same (or the opposite, when using the difference) effect on the phenotype (**Fig. 1a,d**). However, if the effect of the SNPs differs between the two traits, MTMM outperforms these approaches (**Fig. 1b,c**).

It should be noted that, because the background SNPs are correlated due to population structure, simple single-locus tests of association are strongly biased toward false positives, just as in the original data¹⁴. The mixed model effectively removes this bias, regardless of whether we analyze one phenotype at a time using a single-trait mixed model or both simultaneously using MTMM (**Supplementary Fig. 2**). However, analyzing these data with methods that do not take population structure into account is clearly not a realistic option (**Supplementary Fig. 3**).

In addition to the model just described, we simulated an oligogenic scenario in which

each phenotype was determined by 20 loci, each of which could, with equal probability, affect (i) that phenotype only, (ii) both phenotypes in the same way or (iii) both phenotypes but in opposite ways. The behavior of each locus was chosen independently, and the resulting distribution of correlations between the phenotypes was thus centered on zero (**Supplementary Fig. 4**), which is very different from the positively correlated phenotypes generated under the first simulation scenario (**Supplementary Fig. 1**). MTMM is intended for correlated phenotypes and is expected to perform less well when phenotypes are weakly correlated. The oligogenic simulation results supported this intuition. For weakly correlated pairs of phenotypes, single-trait analysis often outperformed MTMM (especially in detecting SNPs with effect in one phenotype only); however, for more strongly correlated phenotypes, the results agreed with those presented above, in that MTMM always outperformed marginal analyses (**Supplementary Fig. 5**). Note that the correlation does not have to be positive: for negatively correlated phenotypes, MTMM has relatively higher power to detect SNPs with the same effect in both phenotypes, whereas for positively correlated phenotypes, it performs best for SNPs that have opposing effects (sometimes it may make sense to simply change the direction of correlation by negating one of the phenotypes when analyzing real data).

As noted, an advantage of MTMM is that it can be used for correlated phenotypes regardless of whether the phenotypes represent different measurements (and the correlations are due to pleiotropy) or the same trait measured in different environments (**Fig. 1a–d**). However, the simulations above assume that the phenotypic correlations are solely due to genetics and not environment, and this is only likely to be true for studies involving inbred lines in controlled environments. Certainly, correlations between pleiotropic traits will reflect environment as well as genotype. To verify that MTMM is able to separate these effects, we simulated another 5,000 pairs of correlated traits using the 10,000-loci model, but now with correlations reflecting environmental as well as genetic covariance (Online Methods). Both the environmental and genetic correlations were well estimated (**Supplementary Fig. 6**), although it should be noted that the residuals of the genetic and environmental correlation estimates

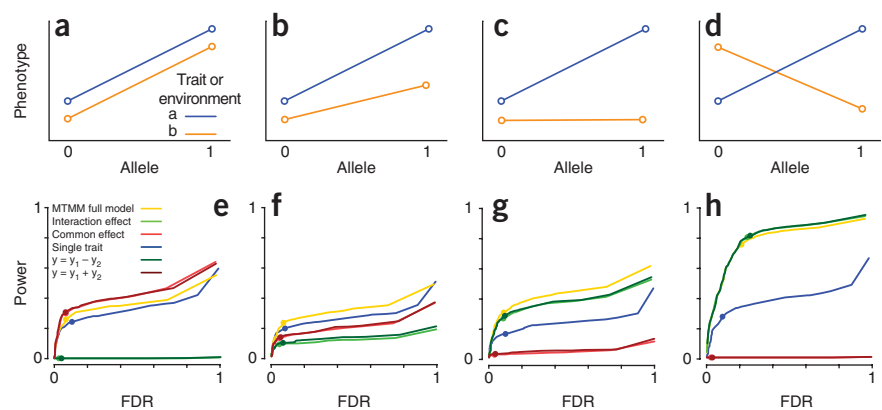


Figure 1 Simulation results. (a–d) Scenarios simulated—with positive pleiotropy, alternative common effect across environments (a); positive pleiotropy, alternative common effect across environments, with size of effect differing between traits and environments (b); effect only on one trait, alternative only in one environment (c); and negative pleiotropy, alternative opposite effect across environments (d). (e) Estimated relationship between power and false discovery rate (FDR) using six different statistical tests for the scenario in a. (f) Estimated relationship between power and FDR for the scenario in b. (g) Estimated relationship between power and FDR for the scenario in c. (h) Estimated relationship between power and FDR for the scenario in d. Dots on curves denote nominal Bonferroni-corrected 5% significance thresholds. Both power and FDR were calculated with respect to the single focal locus only.

Table 1 MTMM estimates of correlation and heritability in NFBC1966 data

	Genetic				Environmental			Heritability ^P
	Phenotypic ^a	Corr.	s.e.m.	P value	Corr.	s.e.m.	P value	
HDL-TG	-0.37	-0.42	0.14	0.024	-0.36	0.06	1.58 × 10 ⁻⁸	0.38/0.18
HDL-LDL	-0.13	-0.19	0.11	0.085	-0.09	0.08	0.26	0.39/0.45
HDL-CRP	-0.19	0.24	0.23	0.25	-0.34	0.06	1.50 × 10 ⁻⁷	0.39/0.14
TG-LDL	0.32	0.31	0.14	0.062	0.35	0.06	9.64 × 10 ⁻⁷	0.19/0.44
TG-CRP	0.21	-0.50	0.39	0.115	0.34	0.05	3.19 × 10 ⁻⁹	0.18/0.13
LDL-CRP	0.09	0.08	0.19	0.65	0.10	0.06	0.12	0.45/0.13

Corr., correlation.

^aDirect estimates of the Pearson correlation are identical to the precision given. ^bThe s.e.m. of all heritability estimates is between 0.05 and 0.06. Single-trait estimates are 0.38 (HDL), 0.18 (triglycerides, TG), 0.45 (LDL) and 0.13 (CRP).

are negatively correlated (Supplementary Fig. 6d). The accuracy of these estimates does affect the performance of GWAS, but the effect seems to be relatively minor (Supplementary Fig. 7).

Pleiotropy in human data

To show the usefulness of MTMM for traits that are correlated because they are part of the same biological system, we reanalyzed data from NFBC1966 (ref. 31) (see Online Methods for details). We focused on measurements of four blood metabolites that are strongly involved in cardiovascular heart disease³³, namely triglycerides, low-density lipoprotein (LDL), high-density lipoprotein (HDL) and C-reactive protein (CRP). These metabolites are significantly correlated, and MTMM analysis indicated that the correlations are caused by genetics as well as environment (Table 1), supporting the notion that these traits are mechanistically related and/or have linked

causal loci. For HDL-CRP and triglycerides-CRP, the correlations of the genetic effects were in the opposite direction of the environmental correlations. However, in these cases, the genetic correlations are not significantly different from zero, and it is likely that the phenotypic correlations are driven primarily by the shared environment.

In terms of associations, the results from the joint analysis of triglycerides and LDL suffice to show how two of our main predictions were borne out. First, almost all SNPs that were found to be significantly associated

in the marginal analysis of either LDL or triglycerides also had significant associations in the joint analysis (Table 2). However, MTMM arguably provides greater insight into the nature of the associations, as it reveals interaction effects. Second, MTMM finds associations that the marginal analyses do not. In particular, for positively correlated phenotypes such as triglyceride and LDL measures, we expect MTMM to have much greater power to detect polymorphisms whose effects differ greatly between the phenotypes. A good example of this is the *FADS1-FADS2* locus, which was not significantly associated in either marginal analysis but showed highly significant association using MTMM because of a very strong interaction effect (Fig. 2 and Table 2). These genes are excellent candidates and were mentioned in the previous analysis of the NFBC1996 data³¹. Notably, they were also identified in a massive meta-analysis involving more than 100,000 individuals³⁴, which furthermore reported opposite effects on

Table 2 SNPs detected in the analysis of LDL and triglycerides using a genome-wide significance of 0.05

SNP	Position	MTMM (P value) ^a			Single-trait mixed model (P value) ^a	
		Full test	Interaction	Common	LDL	TG
<i>CELSR2</i> region, chromosome 1						
rs611917	109616775	6.42 × 10⁻⁸	3.19 × 10 ⁻³	7.72 × 10 ⁻⁷	1.80 × 10⁻⁸	0.46
rs646776	109620053	2.48 × 10⁻¹⁵	1.42 × 10 ⁻⁶	3.28 × 10⁻¹¹	3.92 × 10⁻¹⁵	0.77
<i>APOB</i> region, chromosome 2						
rs10198175	20997364	6.32 × 10 ⁻⁷	0.02	1.33 × 10 ⁻⁶	9.48 × 10⁻⁸	0.29
rs3923037	21011755	6.39 × 10⁻⁹	0.13	2.64 × 10⁻⁹	2.72 × 10 ⁻⁷	7.17 × 10 ⁻⁷
rs6728178	21047434	9.57 × 10⁻¹⁰	0.11	4.37 × 10⁻¹⁰	7.95 × 10⁻⁸	1.81 × 10 ⁻⁷
rs6754295	21059688	1.31 × 10⁻⁹	0.14	4.97 × 10⁻¹⁰	7.10 × 10⁻⁸	4.12 × 10 ⁻⁷
rs676210	21085029	2.43 × 10⁻⁹	0.04	2.56 × 10⁻⁹	7.23 × 10 ⁻⁷	9.21 × 10⁻⁸
rs693	21085700	1.80 × 10⁻¹⁰	0.19	5.00 × 10⁻¹¹	2.84 × 10⁻¹¹	2.79 × 10 ⁻³
rs673548	21091049	1.63 × 10⁻⁹	0.04	1.85 × 10⁻⁹	5.97 × 10 ⁻⁷	6.43 × 10⁻⁸
rs1429974	21154275	4.85 × 10 ⁻⁷	0.02	1.06 × 10 ⁻⁶	7.69 × 10⁻⁸	0.24
rs754524	21165046	5.30 × 10⁻⁸	0.02	1.39 × 10⁻⁷	7.83 × 10⁻⁹	0.17
rs754523	21165196	4.51 × 10 ⁻⁷	0.02	1.01 × 10 ⁻⁶	7.15 × 10⁻⁸	0.24
<i>GCKR</i> region, chromosome 2						
rs1260326	27584444	5.33 × 10⁻¹⁰	2.10 × 10⁻⁸	7.73 × 10 ⁻³	0.21	1.87 × 10⁻¹⁰
rs780094	27594741	5.98 × 10⁻⁹	4.22 × 10⁻⁸	0.01	0.44	3.15 × 10⁻⁹
<i>LPL</i> region, chromosome 8						
rs10096633	19875201	2.42 × 10⁻⁸	2.04 × 10⁻⁸	0.06	0.97	1.93 × 10⁻⁸
<i>FADS1</i> region, chromosome 11						
rs174537	61309256	1.60 × 10⁻⁹	9.02 × 10⁻⁹	0.01	6.82 × 10 ⁻⁶	3.81 × 10 ⁻³
rs102275	61314379	8.79 × 10⁻¹⁰	6.20 × 10⁻⁹	4.86 × 10 ⁻³	4.13 × 10 ⁻⁶	3.82 × 10 ⁻³
rs174546	61326406	5.52 × 10⁻¹⁰	3.83 × 10⁻⁹	4.88 × 10 ⁻³	3.69 × 10 ⁻⁶	3.12 × 10 ⁻³
rs174556	61337211	2.56 × 10⁻⁹	4.43 × 10⁻⁸	1.93 × 10 ⁻³	2.03 × 10 ⁻⁶	0.01
rs1535	61354548	2.08 × 10⁻⁹	1.35 × 10⁻⁸	0.01	6.04 × 10 ⁻⁶	4.96 × 10 ⁻³
rs2072114	61361791	8.77 × 10⁻⁸	7.31 × 10 ⁻⁷	4.77 × 10 ⁻³	1.59 × 10 ⁻⁵	0.03
<i>LDLR</i> region, chromosome 19						
rs11668477	11056030	3.16 × 10⁻⁸	0.15	1.18 × 10⁻⁸	3.89 × 10⁻⁹	0.02
rs2228671	11071912	7.20 × 10⁻⁸	5.30 × 10 ⁻⁴	4.87 × 10 ⁻⁶	4.47 × 10⁻⁸	0.96

^aP values below the Bonferroni-corrected 5% cutoff of 1.5 × 10⁻⁷ are highlighted in bold.

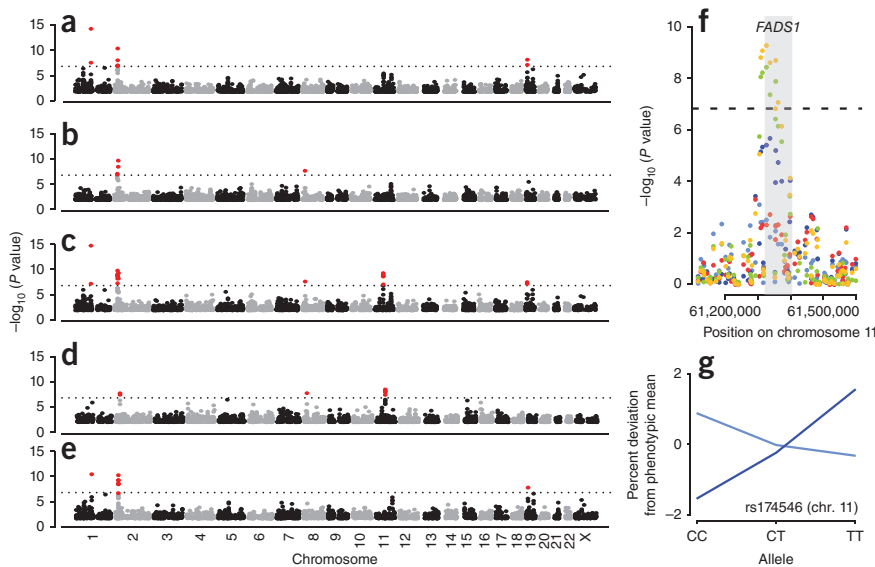


Figure 2 GWAS of LDL and triglycerides. (a,b) Manhattan plots for the marginal, single-trait analyses of LDL (a) and triglycerides (TG) (b). (c–e) Manhattan plots for the joint MTMM analyses with the full model (c), interaction effect (d) and common effect (e). The dotted horizontal lines denote the 5% Bonferroni-adjusted genome-wide significance level. (f) Enlarged view of the *FADS1-FADS2* region on chromosome 11. The points for the single-trait analyses are shown in light blue (TG) and dark blue (LDL), and the point for MTMM are shown in orange (full test), light green (interaction effect) and red (common effect). Gray shading denotes the *FADS1* gene region. (g) Estimated phenotypic effect of the rs174546 SNP is shown in light blue (TG) and dark blue (LDL).

and 7). Perhaps unexpectedly, we found very few interaction effects. Out of a total of 41 significant SNP associations, only 3 seemed to be caused by interactions. A rare allele (minor allele frequency (MAF) = 4%) on chromosome 5

was identified as having a significant genotype-by-season effect, but it does not correspond to any obvious candidate gene (Supplementary Fig. 10). A more convincing example was provided by the two tightly linked and perfectly correlated SNPs on chromosome 1. These were identified by comparing the full model to one without interaction terms, although the interaction with the simulated season seemed to be strongest (Fig. 4). The minor allele (MAF = 3%) was associated with delayed flowering (Fig. 4b), but the effects depend strongly on the season and are much more pronounced in the (simulated) summer. Notably, both SNPs are in the coding region of the *FRS6* gene, which is known to be involved in the phyA-mediated response to far-red light³⁵. Knockout lines of this locus have an early-flowering phenotype, the magnitude of which depends on day length (one of the factors that varies between the simulated seasons).

Of the remaining 38 SNPs, 28 were found by both marginal and joint analysis (as common effects), and 10 were found only by marginal analysis. Although our simulations would seem to suggest that MTMM should always have higher power than marginal tests, even for detecting common or unique effects, this is clearly not always the case. The phenotypes analyzed here are extremely highly correlated as well as heritable (all coefficients are typically well above 0.9; Supplementary Table 6). In such cases, the advantage of increasing the sample size through joint analysis does not necessarily outweigh the cost of a more complex model with more degrees of freedom.

Genotype-environment interactions in *A. thaliana* data

The other natural application for MTMM is when phenotypes are correlated because they represent the same trait measured in different environments. In such a setting, one is often directly interested in finding genes that are involved in the differential response to the environment, that is, genotype-by-environment ($G \times E$) interactions. We tested this application using a data set from *A. thaliana* in which flowering time was measured (for a global collection of naturally occurring inbred lines) in environmental control chambers for two simulated seasons ('spring' and 'summer') and two simulated locations ('Spain' and 'Sweden')³². Flowering time varies in a clinal manner and is generally thought to be important in local adaptation. It is thus both natural and of interest to try identifying genes that are responsible for the differential flowering response to different environments³².

We analyzed the *A. thaliana* data using a full 2×2 factorial model: in addition to estimating the effect of genotype, season and location, we have two pairwise interaction terms (Online Methods and Supplementary Note). The results are summarized in Figure 3 (for details, see Supplementary Figs. 8 and 9 and Supplementary Tables 6

and 7). Perhaps unexpectedly, we found very few interaction effects. Out of a total of 41 significant SNP associations, only 3 seemed to be caused by interactions. A rare allele (minor allele frequency (MAF) = 4%) on chromosome 5

was identified as having a significant genotype-by-season effect, but it does not correspond to any obvious candidate gene (Supplementary Fig. 10). A more convincing example was provided by the two tightly linked and perfectly correlated SNPs on chromosome 1. These were identified by comparing the full model to one without interaction terms, although the interaction with the simulated season seemed to be strongest (Fig. 4). The minor allele (MAF = 3%) was associated with delayed flowering (Fig. 4b), but the effects depend strongly on the season and are much more pronounced in the (simulated) summer. Notably, both SNPs are in the coding region of the *FRS6* gene, which is known to be involved in the phyA-mediated response to far-red light³⁵. Knockout lines of this locus have an early-flowering phenotype, the magnitude of which depends on day length (one of the factors that varies between the simulated seasons).

Of the remaining 38 SNPs, 28 were found by both marginal and joint analysis (as common effects), and 10 were found only by marginal analysis. Although our simulations would seem to suggest that MTMM should always have higher power than marginal tests, even for detecting common or unique effects, this is clearly not always the case. The phenotypes analyzed here are extremely highly correlated as well as heritable (all coefficients are typically well above 0.9; Supplementary Table 6). In such cases, the advantage of increasing the sample size through joint analysis does not necessarily outweigh the cost of a more complex model with more degrees of freedom.

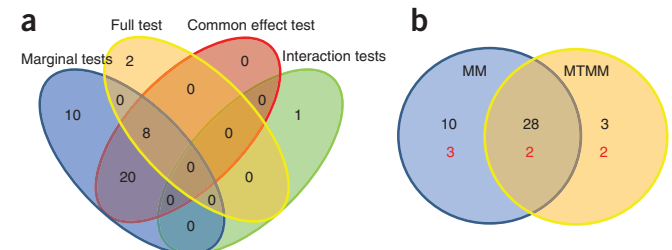
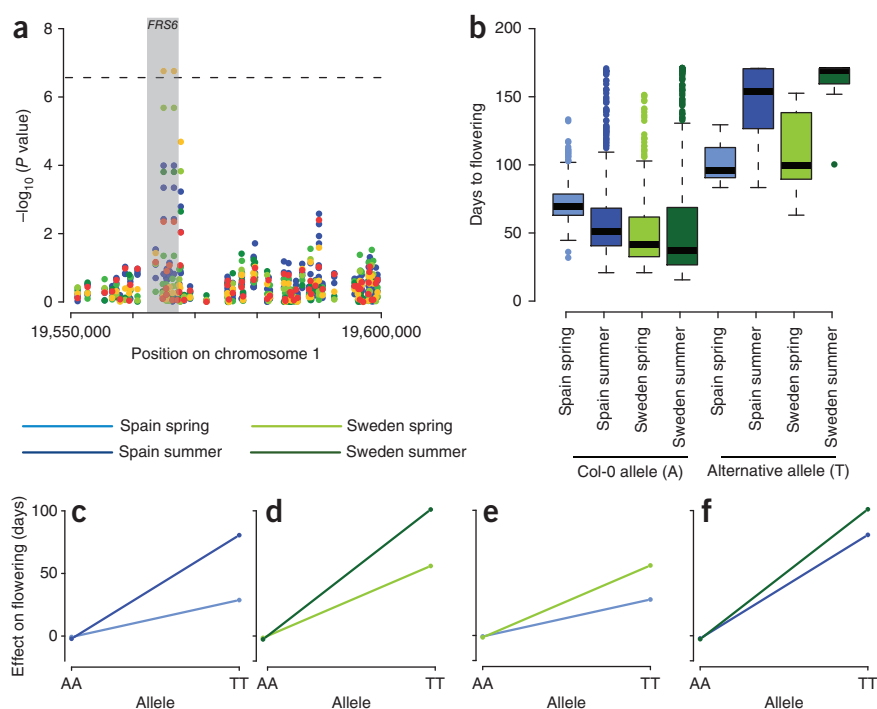


Figure 3 Venn diagrams summarizing the GWAS of *A. thaliana* flowering data³². (a) Classification of the 41 significantly associated SNPs according to the test(s) in which they were significant. (b) Classification of the 41 SNPs (black) and corresponding gene regions (red) according to whether they were found using marginal (mixed model, MM) or joint (MTMM) analysis.

Figure 4 Summary of *FRS6* results. **(a)** Enlarged view of a 50-kb region on chromosome 1 showing significant genotype-environment associations. The *FRS6* gene is highlighted in gray. The results for the four marginal analyses (using a single trait, MM) are shown in blue, and the MTMM results are shown in orange (full test), light green (three-way interaction), green (genotype by location), dark green (genotype by season) and red (common effects). The horizontal line represents the 5% Bonferroni-corrected genome-wide significance threshold. **(b)** Phenotypic distribution shown as box plots, when partitioned by the genotype of the significantly associated SNP in **a** and by the environment. The colored boxes denote the first quantile (bottom), median value (thick black line) and the third quantile (top). The four colors represent the four different environmental conditions. Col-0 represents the *A. thaliana* reference strain. **(c-f)** Plots contrasting the allelic effect in different comparisons: the effect of the season in 'Spain' (**c**), the effect of the season in 'Sweden' (**d**), the effect of the location in 'spring' (**e**) and the effect of the location in 'summer' (**f**). The effect depends strongly on the season within each location (**c,d**) and less strongly on location with season (**e,f**).



DISCUSSION

We have shown how the classical mixed model from breeding may be used for GWAS of correlated phenotypes in structured populations, often providing greater statistical power than marginal analyses. However, we emphasize that our approach is much more than an *ad hoc* method for increasing power. The model we use effectively dates back to Fisher³⁶, and can be derived from basic genetic principles, under the assumption that heritable phenotypic variation is due to very large numbers of genes of very small effect (Online Methods). Assuming that this is a reasonable approximation (and it seems to be, for a growing number of traits), we can disentangle genetic correlations from environmental correlations, whenever these are uncorrelated. This allows us to address fundamental questions about the nature of variation.

When applied to traits that may be biologically related, the resulting variance component estimates allow us to assess the level of pleiotropy without estimating effects of individual loci. Using data from different human blood lipid measures, we demonstrated how the phenotype covariance can be decomposed into genetic and environmental terms, suggesting that most of these traits are indeed correlated due to shared genetics (they are pleiotropic or due to causal sites in linkage disequilibrium). A similar approach was recently used³⁷ to assess the heritability of RNA expression levels within and across human cell tissues.

Irrespective of this, we also showed increased power, detecting several interesting loci affecting human blood lipid level that were not significant in single-trait analysis but that have all been replicated in GWAS using much larger sample sizes. This finding alone strongly argues for routine application of our method to correlated phenotypes.

As an example of how the method can be used to detect environmental interactions, we applied our method to an *A. thaliana* flowering-time data set, where the plants had been phenotyped under four different environmental conditions (in a classical 2 × 2 factorial design). These phenotypes are highly correlated as well as highly heritable, and the estimated variance components suggest that there is in fact very little difference between the environments at the genetic

level (Supplementary Table 6). Hence, it is arguably not unexpected that we detected little in terms of interaction effects. Although it is of course possible that we simply do not have the power to detect interactions, it is notable that analogous studies in maize have also been unable to detect large genome-environment interaction effects³⁸. The results from *A. thaliana* and maize are strikingly different from what has been reported for mouse³⁹, yeast⁴⁰ and even humans⁴, but the reasons for these differences are far from clear, given the dramatically different study designs.

Full factorial designs with replicated genotypes are of course not possible in most organisms; however, we note that MTMM does not require this. Indeed, a mixed-model approach has previously been proposed for estimating genome-environment variance components in humans²⁵ (using a special case of our model in which heritabilities are assumed to be equal across environments; Online Methods). Either approach is directly applicable to human data.

Although we have focused on relatively simple pairwise correlations in this paper, it is easy to model more than two phenotypes using MTMM. Conceptually, we believe that extending this approach to larger multi-trait experiments should allow for greater benefits in estimating error terms and elucidating functional relationships between suites of traits. However, for such complex models, the computational complexity grows fast, and the results become increasingly difficult to interpret compared to sequential two-trait analyses.

This is a well-known problem in statistics and quantitative genetics, but MTMM has the additional caveat that it assumes that the increasingly complex covariance structure, which is estimated in the absence of fixed effects, remains constant as these are added. Various intermediate approaches are possible, for example, variance components might be estimated using a full model once, followed by GWAS using submodels; more work in this area is clearly desirable.

Finally, when the phenotypes are not correlated or if the correlation is not due to genetics (something that can be deduced from the variance component estimates), a single-trait mixed model will generally have greater power to detect causal loci that are phenotype specific. When this will be the case precisely is hard to predict; however, we suggest using

the MTMM approach as a complement to, rather than replacement for, marginal GWAS. The advantages are clear: it allows the detection of both interactions and pleiotropic loci in a rigorous statistical framework while simultaneously accounting for population structure.

URLs. MTMM has been implemented in a set of R scripts (MTMM) for carrying out GWAS. These scripts rely on the software ASREML⁴¹ for the estimation of the variance components. The scripts can be obtained <https://cynin.gmi.oeaw.ac.at/home/resources/mtmm>.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Supplementary information is available in the online version of the paper.

ACKNOWLEDGMENTS

We thank the NFBC1966 Study Investigators for allowing us to use their phenotype and genotype data in our study. The NFBC1966 Study is conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with the Broad Institute, the University of California, Los Angeles (UCLA), the University of Oulu and the National Institute for Health and Welfare in Finland. This manuscript was not prepared in collaboration with investigators of the NFBC1966 Study and does not necessarily reflect the opinions or views of the NFBC1966 Study Investigators, the Broad Institute, UCLA, the University of Oulu, the National Institute for Health and Welfare in Finland or the NHLBI. We furthermore thank N.B. Freimer and S.K. Service for their help in preprocessing the NFBC1966 data. We would also like to thank P. Forai for excellent IT and cluster support at the Gregor Mendel Institute, the INRA MIGALE bioinformatics platform for further computational resources and J. Dekkers, P. Donnelly, E. Eskin, C. Niango and A. Price for comments on the manuscript and/or helpful discussions. This work was supported by grants to M.N. from the US National Institutes of Health (P50 HG002790) and the European Union Framework Programme 7 (TransPLANT, grant agreement 283496), as well as by grants from the Deutsche Forschungsgemeinschaft (DFG) (A.K., KO4184/1-1) and the Ecologie des Forêts, Prairies et milieux Aquatiques (EFPA) department of INRA (V.S.).

AUTHOR CONTRIBUTIONS

All authors helped design the study. A.K., B.J.V. and V.S. developed the theory and implemented the simulations. A.K., B.J.V. and M.N. wrote the paper with input from V.S., A.P. and Q.L.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at <http://www.nature.com/doi/10.1038/ng.2376>.

Reprints and permissions information is available online at <http://www.nature.com/reprints/index.html>.

- Platt, A., Vilhjálmsson, B.J. & Nordborg, M. Conditions under which genome-wide association studies will be positively misleading. *Genetics* **186**, 1045–1052 (2010).
- Kang, H.M. *et al.* Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* **42**, 348–354 (2010).
- Price, A.L., Zaitlen, N.A., Reich, D. & Patterson, N. New approaches to population stratification in genome-wide association studies. *Nat. Rev. Genet.* **11**, 459–463 (2010).
- Hamza, T.H. *et al.* Genome-wide gene environment study identifies glutamate receptor gene *GRIN2A* as a Parkinson's disease modifier gene via interaction with coffee. *PLoS Genet.* **7**, e1002237 (2011).
- Lynch, M. & Walsh, B. *Genetics and Analysis of Quantitative Traits* (Sinauer Associates, Sunderland, Massachusetts, 1997).
- Jiang, C. & Zeng, Z.B. Multiple trait analysis of genetic mapping for quantitative trait loci. *Genetics* **140**, 1111–1127 (1995).
- Ferreira, M.A. & Purcell, S.M. A multivariate test of association. *Bioinformatics* **25**, 132–133 (2009).
- Zhang, L., Pei, Y.F., Li, J., Papasian, C.J. & Deng, H.W. Univariate/multivariate genome-wide association scans using data from families and unrelated samples. *PLoS ONE* **4**, e6502 (2009).
- Knott, S.A. & Haley, C.S. Multitrait least squares for quantitative trait loci detection. *Genetics* **156**, 899–911 (2000).
- Henderson, C.R. *Application of Linear Models in Animal Breeding* (University of Guelph, Guelph, Canada, 1984).
- Thomas, D. Gene-environment-wide association studies: emerging approaches. *Nat. Rev. Genet.* **11**, 259–272 (2010).
- Ober, C. & Vercelli, D. Gene-environment interactions in human disease: nuisance or opportunity? *Trends Genet.* **27**, 107–115 (2011).
- Yu, J. *et al.* A unified mixed model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.* **38**, 203–208 (2006).
- Atwell, S. *et al.* Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. *Nature* **465**, 627–631 (2010).
- Huang, X. *et al.* Genome-wide association studies of 14 agronomic traits in rice landraces. *Nat. Genet.* **42**, 961–967 (2010).
- Olsen, H.G. *et al.* Genome-wide association mapping in Norwegian Red cattle identifies quantitative trait loci for fertility and milk production on BTA12. *Anim. Genet.* **42**, 466–474 (2011).
- Tian, F. *et al.* Genome-wide association study of leaf architecture in the maize nested association mapping population. *Nat. Genet.* **43**, 159–162 (2011).
- Zhao, K. *et al.* An *Arabidopsis* example of association mapping in structured samples. *PLoS Genet.* **3**, e4 (2007).
- Kang, H.M. *et al.* Efficient control of population structure in model organism association mapping. *Genetics* **178**, 1709–1723 (2008).
- Zhang, Z. *et al.* Mixed linear model approach adapted for genome-wide association studies. *Nat. Genet.* **42**, 355–360 (2010).
- Idaghdour, Y. *et al.* Geographical genomics of human leukocyte gene expression variation in southern Morocco. *Nat. Genet.* **42**, 62–67 (2010).
- International Multiple Sclerosis Genetics Consortium and Wellcome Trust Case Control Consortium 2. Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature* **476**, 214–219 (2011).
- Stich, B., Piepho, H.P., Schulz, B. & Melchinger, A.E. Multitrait association mapping in sugar beet (*Beta vulgaris* L.). *Theor. Appl. Genet.* **117**, 947–954 (2008).
- Lee, S.H., Wray, N.R., Goddard, M.E. & Visscher, P.M. Estimating missing heritability for disease from genome-wide association studies. *Am. J. Hum. Genet.* **88**, 294–305 (2011).
- Yang, J., Lee, S.H., Goddard, M.E. & Visscher, P.M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).
- Lee, S.H. *et al.* Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. *Nat. Genet.* **44**, 247–250 (2012).
- Deary, I.J. Genetic contributions to stability and change in intelligence from childhood to old age. *Nature* **482**, 212–215 (2012).
- Kim, S. & Xing, E.P. Statistical estimation of correlated genome associations to a quantitative trait network. *PLoS Genet.* **5**, e1000587 (2009).
- Manning, A.K. *et al.* Meta-analysis of gene-environment interaction: joint estimation of SNP and SNP × environment regression coefficients. *Genet. Epidemiol.* **35**, 11–18 (2011).
- Horton, M.W. *et al.* Genome-wide patterns of genetic variation in worldwide *Arabidopsis thaliana* accessions from the RegMap panel. *Nat. Genet.* **44**, 212–216 (2012).
- Sabatti, C. *et al.* Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nat. Genet.* **41**, 35–46 (2009).
- Li, Y., Huang, Y., Bergelson, J., Nordborg, M. & Borevitz, J.O. Association mapping of local climate-sensitive quantitative trait loci in *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci. USA* **107**, 21199–21204 (2010).
- Kathiresan, S. *et al.* A genome-wide association study for blood lipid phenotypes in the Framingham Heart Study. *BMC Med. Genet.* **8** (suppl. 1) S17 (2007).
- Teslovich, T.M. *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**, 707–713 (2010).
- Lin, R. & Wang, H. *Arabidopsis FHY3/FAR1* gene family and distinct roles of its members in light control of *Arabidopsis* development. *Plant Physiol.* **136**, 4010–4022 (2004).
- Fisher, R. The correlation between relatives on the supposition of Mendelian inheritance. *Trans. R. Soc. Edinburgh* **52**, 399–433 (1918).
- Price, A.L. *et al.* Single-tissue and cross-tissue heritability of gene expression via identity-by-descent in related or unrelated individuals. *PLoS Genet.* **7**, e1001317 (2011).
- Buckler, E.S. *et al.* The genetic architecture of maize flowering time. *Science* **325**, 714–718 (2009).
- Valdar, W. *et al.* Genetic and environmental effects on complex traits in mice. *Genetics* **174**, 959–984 (2006).
- Smith, E.N. & Kruglyak, L. Gene-environment interaction in yeast gene expression. *PLoS Biol.* **6**, e83 (2008).
- Gilmour, A., Gogel, B., Cullis, B., Welham, S.J. & Thompson, R. *ASReml User Guide Release 1.0* (VSN International, Hemel Hempstead, UK, 2002).

ONLINE METHODS

Theory Multiple-traits mixed model. Following Henderson¹⁰, we can write the mixed model for the phenotypes of n individuals as

$$\mathbf{y} = X\boldsymbol{\beta} + \mathbf{g} + \mathbf{e} \quad (1)$$

where \mathbf{y} is a vector of n phenotype values. In this notation, the trait mean is included, together with other fixed effects, in the design matrix X . $\boldsymbol{\beta}$ represents the effect size of the fixed effects, $\mathbf{g} \sim N(\mathbf{0}, \sigma_g^2 K)$ is a random effect and $\mathbf{e} \sim N(\mathbf{0}, \sigma_e^2 I)$. It follows that the covariance matrix for the trait values \mathbf{y} is

$$\text{var}(\mathbf{y}) = \sigma_g^2 K + \sigma_e^2 I \quad (2)$$

where K is an $n \times n$ kinship or relatedness matrix. If we consider two traits, \mathbf{y}_1 and \mathbf{y}_2 , measured on the same set of individuals, then under the mixed model for the k th phenotype follows the partitions of the variance accordingly: $\text{var}(\mathbf{y}_k) = \sigma_{gk}^2 K + \sigma_{ek}^2 I$. However, for the covariance matrix between the two phenotypes, it is not obvious what the appropriate model is. Henderson⁴² suggests the covariance model

$$\text{cov}(\mathbf{y}_1, \mathbf{y}_2) = \sigma_{g1}\sigma_{g2}\rho_g K + \sigma_{e1}\sigma_{e2}\rho_e I \quad (3)$$

Where ρ_g captures the genetic correlation between two phenotypes and the term ρ_e captures the correlation caused by shared environment and other nongenetic sources of correlation.

We can generalize this for phenotypes that have been measured for different sets of individuals (**Supplementary Note**).

Estimating the variance parameters. The estimation procedure for the variance components is described in the **Supplementary Note**.

Application to GWAS. As in EMMAX² or P3D²⁰, we estimate the covariance matrix only once to re-estimate a scalar in front of it for every marker. This fixes five degrees of freedom out of six in total (maximum number of variance components for two traits). For a pair of traits (the i th and j th traits), the proposed approximation effectively assumes that the three variance ratios (σ_{gi}/σ_{ei} , σ_{gj}/σ_{ej} and σ_{gi}/σ_{gj}) and the two correlations ρ_{gi} and ρ_{gj} are fixed with and without the marker in the model.

With multiple traits, we can search for causal loci with common effects (across all traits) as well as trait-specific loci or loci with opposite effects for different traits. Depending on what we are interested in, a generalized least squares (GLS) F test can be constructed to compare two models. For two traits, we can write the single marker model as

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \mathbf{s}_1\mu_1 + \mathbf{s}_2\mu_2 + \mathbf{x}\boldsymbol{\beta} + (\mathbf{x} \times \mathbf{s}_1)\boldsymbol{\alpha} + \mathbf{v} \quad (4)$$

where \mathbf{x} is the marker and \mathbf{s}_i is a vector of 1 for all values belonging to the i th trait and 0 otherwise. $\mathbf{v} \sim N(\mathbf{0}, \text{cov}(\mathbf{y}))$ is a random variable capturing both the error and genetic random effects. Depending on what kind of loci we are interested in, we propose three different F tests.

1. The full model tested against a null model where $\boldsymbol{\beta} = \mathbf{0}$ and $\boldsymbol{\alpha} = \mathbf{0}$. This identifies both loci with common and differing effects in one model but suffers in power from the extra degree of freedom.
2. To identify common genetic effects, we propose to test the genetic model ($\boldsymbol{\alpha} = \mathbf{0}$) against a null model where $\boldsymbol{\beta} = \mathbf{0}$ and $\boldsymbol{\alpha} = \mathbf{0}$.
3. Finally, to identify differing genetic effects between the traits, we propose to test the full model against a null model where $\boldsymbol{\alpha} = \mathbf{0}$.

As both the interaction test and the common effect test are sensitive to scaling of the phenotype values, we propose to normalize them either by the total variance or the genetic variance (as obtained in marginal trait analysis). To minimize multiple-testing problems, one could, for example, carry out GWAS using the full model and then use the other tests to analyze associated loci further.

Extending this model for an arbitrary number of traits is straightforward (one example for the analysis of four traits is described in the **Supplementary Note**). However, when there are more than two traits in the model, the number of possible tests grows quickly. A noteworthy special case is when there are several environmental variables in a factorial study design, in which case each environmental variable can be included in the model instead of the term

$$\sum_{i=1}^t \mathbf{s}_i \mu_i$$

and their interactions with the genotype could replace the term

$$\mathbf{x} \left(\sum_{i=1}^t \mathbf{s}_i \boldsymbol{\alpha} \right)$$

This can result in a simpler and a more tractable model than if all possible combinations of environments were treated as independent.

Genotype-environment interactions. Given two measured phenotype vectors, \mathbf{y}_1 and \mathbf{y}_2 , Yang *et al.*²⁵ include a $G \times E$ random effect in a mixed model as

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = X\boldsymbol{\beta} + u_G + u_{G \times E} + \mathbf{e} \quad (5)$$

Where μ_G and $\mu_{G \times E}$ are random effects and have covariance matrices as follows:

$$\begin{aligned} \text{cov} \left(\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} \right) &= \text{cov}(u_G) + \text{cov}(u_{G \times E}) + \text{cov}(\mathbf{e}) \\ &= \sigma_G^2 \begin{bmatrix} K_{11} & K_{21} \\ K_{12} & K_{22} \end{bmatrix} + \sigma_{G \times E}^2 \begin{bmatrix} K_{11} & 0 \\ 0 & K_{22} \end{bmatrix} + I \end{aligned} \quad (6)$$

Compared to the model proposed in equation (4), this model implicitly assumes two things: (i) that there are no environmental correlations and (ii) that the heritabilities are the same in each environment, such that $h_1^2 = h_2^2$. As the individuals are different in each environment, the first assumption is appropriate. However, the second assumption is not guaranteed to hold in general, and we therefore propose relaxing it.

Simulations. 10,000-loci model. We simulated 2,000 pairs of correlated phenotypes using a model under which the phenotypes consisted of one randomly chosen SNP with a 'large' (additive) effect, accounting for up to 2% of the total phenotypic variance, and 10,000 randomly chosen SNPs with small additive effects. The effects sizes were drawn from a normal distribution, and a random error was added to fix the trait heritability to 0.95. To ensure variation in trait correlations, all trait pairs shared a random fraction of the 10,000 causal loci, with the fraction drawn from a uniform distribution. The four phenotypic models were distinguished by different effect correlations at the major locus (**Fig. 1**).

In addition, we simulated 5,000 pairs of correlated traits with environmental correlations. We fixed the heritability to 0.5 and allowed the genetic correlation to vary from -1 to 1. Additionally, we added a shared environmental term to the model, mimicking scenarios for both negative and positive environmental correlation.

20-loci model. We also simulated 1,000 pairs of correlated phenotypes using a 20-loci model. Each phenotype was determined by 20 SNPs, where each SNP was randomly assigned to one of the following three categories with equal probabilities: (i) SNPs with same effect in both phenotypes, (ii) SNPs with opposite effect in the two phenotypes and (iii) SNPs with effect in one trait only. The SNPs had additive effects drawn from an exponential distribution. Finally, a random error was added to fix the heritabilities to 0.95. To obtain a single P value for two traits, the smaller of the two P values for each SNP from the marginal mixed model analysis was retained.

Power calculations. For calculation of the power and FDR, any significantly associated SNP within 50 kb of any (or the) causal SNP was classified as a true positive; otherwise, it was classified as a false positive. The results were almost independent of the window size used (**Supplementary Fig. 11**). More important is the effect of the causal SNP(s). The nearly twofold increase of power observed at an FDR of 0.1 in **Figure 1** depended on the effect size of the simulated SNP (**Supplementary Fig. 12**). Throughout this paper, we used the single-analysis Bonferroni-corrected 5% significance threshold.

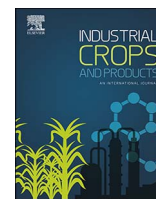
Human data. We used results from the NFBC1966, which consist of phenotypic and genotypic data for 5,402 individuals³¹. Using the exact same data set as was used in ref. 2, after filtering, the data set consisted of

5,326 individuals and 331,475 SNPs. To expedite mapping, unknown genotypes (<1% of the data set) were imputed by replacing missing values with the average genotypic value. Neither the marginal mixed model analysis nor the MTMM tests showed evidence of confounding due to population structure (**Supplementary Fig. 13**).

Analysis of A. thaliana data. The genotype data for *A. thaliana* consisted of 1,307 individuals genotyped at 214,051 SNPs using a custom Affymetrix SNP chip³⁰. The phenotypes used were measurements of flowering time for 459

accessions³². Flowering time was measured in plants grown in four different environments, a factorial setting with two simulated seasons (spring and summer) and two simulated locations (Spain and Sweden). Analyzing the four phenotype vectors together, we can derive five different *F* tests (**Supplementary Note**). None of these tests showed evidence of confounding due to population structure (**Supplementary Fig. 14**).

42. Henderson, C. & Quaas, R.L. Multiple trait evaluation using relatives' records. *J. Anim. Sci.* **43**, 1188–1197 (1976).



Research paper

Near-infrared spectroscopy enables the genetic analysis of chemical properties in a large set of wood samples from *Populus nigra* (L.) natural populations



Mesfin Nigusie Gebreselassie^a, Kévin Ader^{a,b}, Nathalie Boizot^{a,b}, Frédéric Millier^{a,b}, Jean-Paul Charpentier^{a,b}, Ana Alves^c, Rita Simões^c, José Carlos Rodrigues^c, Guillaume Bodineau^d, Francesco Fabbrini^{e,f}, Maurizio Sabatti^e, Catherine Bastien^a, Vincent Segura^{a,*}

^a INRA, UR588 Amélioration, Génétique et Physiologie Forestières, Orléans, France

^b INRA, Plateforme régionale Génobois, Orléans, France

^c Centro de Estudos Florestais, Instituto Superior de Agronomia, 1349-017 Lisboa, Portugal

^d INRA, UE995 Génétique et Biomasse Forestières, Orléans, France

^e Department for Innovation in Biological, Agro-food and Forest systems, University of Tuscia, 01100 Viterbo, Italy

^f Alasia Franco Vivai s.s., Strada Solerette 5/A, 12038 Savigliano, Italy

ARTICLE INFO

Keywords:

Populus nigra L.

Near-infrared spectroscopy

Cell wall composition

Genetic variation

Clonal repeatability

Genotype × Environment interaction

ABSTRACT

High-throughput techniques for the compositional analysis of lignocellulosic biomass are essential to allow the genetic analysis and genetic improvement of bioenergy feedstocks. In this study, we investigated the feasibility of using near-infrared (NIR) spectroscopy for rapid assessment of wood chemical traits in a large sample of *Populus nigra* L. individuals evaluated in clonal trials at two contrasting sites. Spectra were acquired from 5799 wood samples collected in 3 harvests corresponding to two coppice rotations at one site and one coppice rotation at the second. Calibrations were developed and validated using 120 reference samples, representing spectral and chemical variations in the samples. The resulting global and site specific calibrations for most of the traits were at least good enough for ranking of genotypes, demonstrating the usefulness of NIR analysis for phenotyping the studied population. Clonal repeatability (H_c^2) estimates of the studied traits based on all samples were moderate to high (H_c^2 ranging from 0.57 to 0.89 in the 3 harvests). When data were pooled over coppice rotations or sites, the genotype × environment interaction was more evident across sites than across rotations. However, the interaction was smaller than the genotype main effect for all traits, except for glucose and extractives contents. Importantly, the interaction resulted mainly from re-ranking of a few genotypes leaving a substantial amount of stable and performant genetic material, which may encourage breeding for improved main wood components. Optimization of the NIR analysis for assessing clonal trials would facilitate the exploitation of standing genetic variation of energy or chemical related traits in tree breeding program.

1. Introduction

There is currently a considerable interest in moving to alternative and sustainable sources of energy because of the increasing global energy demand, depletion of fossil fuel reserves, fossil fuel-derived climate change and energy related geopolitical tensions. To circumvent some of the prevailing challenges, special focus has recently been given to the production of biofuels from lignocellulosic biomass. Lignocellulosic ethanol is expected to provide a large share of global transportation fuel needs with much less adverse effects than fossil fuels (Schubert, 2006; Sticklen, 2008). However, realizing this potential will

require the synchronized occurrence of genetically improved material, suitable biomass production systems and bioconversion technologies that efficiently convert biomass into bioethanol (Ragauskas et al., 2006; Rubin, 2008).

Candidate biomass feedstocks for the production of second generation bioethanol comprise perennial grasses (e.g., switchgrass and *Miscanthus*) and forest trees (e.g., poplars, *Eucalyptus* and willow) (Abramson et al., 2010). Comparative advantages of poplars (*Populus* spp. and hybrids) in the impending green economy include their rapid growth rates (Bradshaw et al., 2000), good coppicing ability (Ceulemans and Deraedt, 1999) and favourable cell wall chemistry

* Corresponding author.

E-mail address: vincent.segura@inra.fr (V. Segura).

(Guerra et al., 2013; Porth et al., 2013; Wegrzyn et al., 2010). In particular, *Populus nigra* possesses many important characteristics such as adaptability, rooting ability of stem cuttings and resistance to diseases that make it attractive as parent in several hybrid breeding programs in Europe (Cagelli and Lefevre, 1995; Frison et al., 1994).

The major source of lignocellulosic biomass is the plant cell wall, a heterogeneous complex mainly composed of cellulose, hemicelluloses and lignin, with the cellulose microfibrils and the hemicellulosic chains being embedded in lignin (Rubin, 2008). For bioethanol production, the polysaccharides (cellulose and hemicelluloses) are of particular interest because their enzymatic hydrolyses release fermentable monomeric sugars during saccharification. Poplars show substantial variability in cell wall composition, with cellulose content ranging from 42 to 49%, hemicellulose from 16 to 23%, and lignin from 21 to 29% (Sannigrahi et al., 2010). More recently, substantial genetic variation in cell wall chemical traits has been reported for black cottonwood (*P. trichocarpa*) (Guerra et al., 2016; Porth et al., 2013; Wegrzyn et al., 2010) and black poplar (*P. nigra*) (Guerra et al., 2013).

A critical bottleneck in efficient and cost-effective biomass saccharification for bioethanol production is the natural recalcitrance of plant cell walls to enzymatic hydrolysis (Rubin et al., 2007). The most obvious way to reduce biomass recalcitrance is through genetic improvement of trees for wood chemical composition. Poplar breeding for bioenergy can take advantage of past improvements in growth and disease resistance. However, current poplar clonal varieties have not been selected and bred for the qualitative characteristics of the biomass. Thus, there is a need to explore the potential for improvement of cell wall composition to release fermentable sugars and subsequently integrate biorefinery related selection criteria into poplar tree breeding programs. More specifically, development of dedicated bioenergy poplar for future biorefineries requires an understanding of the genetic architecture (extent of genetic variation and covariation, degree of genetic control, underlying polymorphisms/alleles) of both biomass production and biomass composition. This, in turn, accelerates the selection or development of new clones that produce high biomass yields, which are more amenable to bioconversion.

A recent approach to dissect the genetic architecture of “hard-to-measure” complex traits, such as lignocellulosic biomass quality, is to combine high-throughput phenotyping and genomics (Yang et al., 2014). The discovery and analysis of genetic information have been facilitated by the advances in high-throughput sequencing and genotyping platforms together with the availability of reference genome sequences for model forest tree species (Neale and Kremer, 2011). However, high-throughput phenotyping is lagging behind genomics (Araus and Cairns, 2014). Standard methods, such as wet chemistry, used for assessing the chemical composition of wood are costly and low-throughput, which limit their use for assaying of large number of samples as required in genetic studies and breeding programs. As a consequence, the genetic analysis and genetic improvement of cell wall composition may be hindered.

Near-infrared (NIR) spectroscopy is a high-throughput technology that can be applied towards the rapid characterization of a large number of lignocellulosic biomass samples with minimal cost. It is an indirect method based on multivariate statistical analysis to establish relationship between NIR absorbance spectra and reference values of properties of interest using a representative sample set. NIR spectroscopy has been successfully used to predict wood chemical traits in many forest tree species (Tsuchikawa and Kobori, 2015), including *Populus* (Robinson and Mansfield, 2009; Zhang et al., 2014; Zhou et al., 2011), *Eucalyptus* (Alves et al., 2012, 2011; Baillères et al., 2002; Poke and Raymond, 2006; Raymond and Schimleck, 2002), and *Pinus* (Alves et al., 2006; Jiang et al., 2014; Schwanninger et al., 2011a,b; Schwanninger and Hinterstoisser, 2011). Indeed, some studies have utilized NIR predictions for estimating genetic parameters of wood properties, mainly in *Pinus* (Da Silva Perez et al., 2007; Gaspar et al., 2011; Isik et al., 2011) and *Eucalyptus* (Costa e Silva et al., 2008;

Hamilton et al., 2009; Kube et al., 2001; Poke et al., 2006; Raymond et al., 2001; Raymond and Schimleck, 2002; Schimleck et al., 2004; Stackpole et al., 2011, 2010).

To our knowledge, the evaluation of calibration models covering standing genetic variation available in natural and breeding populations of poplar is limited. NIR calibration is useful in genetic studies and selection/breeding activities because such applications require assessment of phenotypes in a large number of samples collected in multi-site environments. In this context, development of calibrations mainly depends on the range of variation of the traits of interest within and across environments. For poplars, this range may be defined not only by the genetic composition of the study population but also by the environmental conditions of the plantation site, short rotation coppice (SRC) management and the age of the tree at sampling time. The purpose of this study was to develop NIR calibration models to predict wood chemical properties, with the aim of applying the predictions to evaluate their genetic variability in natural populations of European black poplar covering the range of the species in Western Europe. Also, the resulting calibrations could be used for rapid screening of elite *P. nigra* clones from natural populations to be used in breeding programs. More specifically, this paper addresses the following objectives: (1) to develop and evaluate calibrations for predicting phenotypes of wood chemical traits in a large sample size ($n = 5799$) based on NIR spectra, (2) to estimate genetic variation in wood chemical properties of young trees and the degree of their genetic control, and (3) to quantify the magnitude and investigate the nature of genotype \times environment ($G \times E$) interaction of the same traits measured across coppice rotations as well as across sites.

2. Materials and methods

2.1. Wood samples and sample preparation

Clonally replicated trials of a *P. nigra* association population were established in 2008 at two contrasting sites located in central France (Orléans, ORL) and northern Italy (Savigliano, SAV) under a SRC system. At each site, a randomized complete block design (RCBD) was used, with a single tree per block and six replicates per genotype. The *P. nigra* population assayed in this study represent the natural range of the species in Western Europe, as it was composed of a diverse set of 1160 cloned genotypes (hereafter, each cloned genotype referred to as genotype) sampled in 14 natural metapopulations across 11 river catchments of four European countries (Table 1). More details concerning the experimental design, site characteristics (soil, climate) and plantation management practices can be found in Guet et al. (2015).

For the analysis of wood chemical properties, a total of 5799 wood samples were taken at 1 m above the ground from 2-yr-old trees in three different harvests (rotations/sites): (i) 289 genotypes in 3 blocks resulting in 795 samples harvested in ORL in March 2010 (end of first coppice cycle, 2008–2009) (hereafter referred to as ORL2010); (ii) 1066 genotypes in 3 blocks resulting in 2805 samples harvested in ORL in February 2012 (end of second coppice cycle, 2010–2011) (hereafter referred to as ORL2012); and (iii) 777 genotypes in 3 blocks resulting in 2199 samples harvested in SAV in January 2011 (end of second coppice cycle, 2009–2010) (hereafter referred to as SAV2011). Circumference at 1 m was measured on all trees of the two sites just before harvest. For each harvest, the final number of biological replicates per genotype ranged between 2 and 3 because of mortality. The samples collected in ORL in 2010 and 2012 have been harvested during two successive 2-yr rotations of the same stool. The wood samples were oven dried at 30 °C for several days until a constant weight was reached, shredded into small pieces with a big cutter and milled using RETSCH SM 2000 cutting mills (SM2000, Retsch, Haan, Germany) to pass through a 1 mm metal sieve in order to get biomass powders onto which NIR spectra were collected. The wood samples were not debarked and both NIR measurements and biochemical analysis were made on non-debarked wood samples.

Table 1

Location, river management and number of studied genotypes in ORL and SAV for the 14 *P. nigra* metapopulations. Where metapopulations were represented by individual trees sampled in different stands distributed along one river, a range of latitudes, longitudes and altitudes is given. Metapopulations were ordered by country according to the latitude of origin. Altitude is expressed in metres a.s.l.

Country	River catchment	Metapopulation	Latitude	Longitude	Altitude	Cohorts ^a	River management ^b	Number of studied genotypes		
								ORL	SAV	Common
France	Adour	Adour	42°53'N–43°23'N	0°02'W–00°56'W	52–902	Mature	Partially regulated	62	52	49
Italy	Basento	Basento	40°24'N – 40°38'N	15°56'E – 16°39'E	37–286	Juvenile/ mature	Partially regulated	26	15	14
France	Dranse	Dranse	46°23'N	06°30'E	374	Juvenile/ mature	Dynamic	40	42	39
France	Durance	Durance	43°51'N	04°59'E	60	Juvenile/ mature	Partially regulated	14	8	1
Germany	Kuhkopf	Kuhkopf	49°49'N	08°30'E	91	Juvenile/ mature	Regulated	53	46	37
France	Loire	Loire	47°00'N – 47°51'N	00°44'W – 02°58'E	29–154	Juvenile/ mature	Dynamic	215	197	165
Netherlands	NL	NL	50°31'N – 52°37'N	03°35'E – 06°23'E	0 – 287	Mature	Regulated	47	42	37
France	Nohèdes	Nohede	42°37'N	02°17'E	820	Mature	Dynamic	43	38	35
Italy	Paglia	Paglia	42°45'N–42°52'N	11°45'E–11° 55'E	235–358	Juvenile/ mature	Dynamic	47	42	41
France	Drôme	Ramieres	44°41'N–44°45'N	04°55'E–05°24'E	145	Juvenile/ mature	Dynamic	178	99	91
France	Rhin	Rhin	48°16'N–48°37'N	07°41'E–07°49'E	135–160	Mature	Regulated	66	50	48
Italy	Stura	Stura	44°17'N – 44°23'N	06°56'E – 07°12'E	825–1699	Juvenile/ mature	Dynamic	25	29	25
Italy	Ticino	Ticino	45°12'N–45°16'N	08°59'E–09°04'E	60–70	Juvenile/ mature	Dynamic	103	78	62
France	Allier	ValAllier	46°24'N	03°19'E	220	Juvenile/ mature	Dynamic	147	39	39

^a Juvenile trees were defined as non-reproductive trees.

^b Regulated if water flows have been regulated to facilitate navigation or to prevent floods; dynamic if water flows are not regulated and allow some flooding events.

2.2. NIR spectra collection, pretreatment and selection of reference samples

Once established, NIR calibration models can be an inexpensive and high-throughput method for accurate estimation of wood chemical properties. However, their initial development involves several steps, including spectral data collection, spectral data pretreatment, selection and analysis of reference samples, application of a multivariate calibration method, model selection and model validation. The NIR spectra of 5799 wood powder samples were measured with a spectrometer Spectrum 400 (Perkin Elmer, Waltham, MA, USA) over 45 days between the end of April 2015 and the beginning of July 2015. Prior to analysis, samples were stabilized in a climatized chamber (20%RH) at 24 °C for a minimum of 1 day. Samples in quartz cups were placed in a rotating device above the integration sphere window and spectra acquired in a temperature controlled room (24 °C). All measurements were done in diffuse-reflectance mode and the obtained spectra were computed as Log (1/R) and expressed in absorbance. The scanning range for all samples was from 10,000 cm⁻¹ to 4000 cm⁻¹ (1000–2500 nm) with a spectral resolution of 8 cm⁻¹ and a zero filling factor of 4 resulting in a number of data points at every 2 cm⁻¹. For each wood sample, 64 scans were acquired and averaged. Background was carried out regularly using Spectralon® as reference.

Undesirable sources that likely affect the quality of spectral data include sample moisture content, particle size, temperature and humidity of the spectrometer laboratory, batch effects (e.g., date of spectral data collection) and so on. Before applying multivariate analysis methods such as partial least squares (PLS) regression, it is important to reduce or remove undesired variations in the recorded sample spectra to reduce noise and enhance calibrations. For this reason, several common spectral pretreatment techniques (normalization, detrend, first and second derivatives on raw or normalized spectra) were applied to the raw spectra for comparisons or to find the best combination. Absorption spectra were first restricted to the wavenumber range of 8000–4000 cm⁻¹ since spectra recorded within 10,000–8000 cm⁻¹ has mainly noise. For illustration, plot of raw

spectra of wood powder samples from ORL2012 harvest is shown in Fig. S1. The spectra pretreatment was performed with R software (R Core Team, 2015). The R packages *prospectr* (Stevens and Ramirez-Lopez, 2013) and *signal* (signal developers, 2013) were used to perform detrend and derivations, respectively. Since spectral data pretreatment can improve exploratory analysis, principal component analysis (PCA) was performed on the resulting 7 spectra modalities (raw, pretreated) to explore the data for potential outlying spectra and clustering of the samples according to genotypes, date of spectral data collection, temperature and humidity of the spectrometer laboratory and operators (not shown). In this initial exploratory analysis, no samples were removed as outliers.

Careful selection of representative samples (reference samples) is a prerequisite to develop NIR spectra based calibrations. We chose to select 120 reference samples based on spectral data because the NIR spectra basically contains information about several properties of a wood sample for which the calibration is carried out. These samples should therefore be selected in order to represent most of the spectral variation of a large population of wood samples (n = 5799) collected from a multi-environment experiment. Also, the reference samples should best represent the sources of variation likely to occur in future samples such as plantation site, coppice rotation and genotype, which could enhance the robustness of the resulting calibrations. To do so, we first calculated the mean spectrum for each genotype within harvest. PCA was then performed on the resulting genotypic spectra across all harvests. The results obtained provided two types of information. First, compared to other spectra modalities, first derivative spectra (first derivative on raw spectra) showed a more uniform distribution of the genotypic spectra on the first 2 PCs (Fig. S2) and were thus chosen to be used for the selection of reference samples. Second, the genotypic spectra showed clear clusters in the space of the first 2 PCs according to harvests (Fig. S2). We thus decided to select an equal number of genotypes from each harvest to constitute the reference sample set. Euclidean distances were computed between the genotypic spectra within harvests. Subsequently, a representative subset of genotypes was

selected within each harvest following the Kennard-Stone algorithm which allows to select samples with a uniform distribution over the predictor space (Kennard and Stone, 1969). A total of 45 genotypes (i.e., 14–16 genotypes per harvest) were selected in order to reach a total of 120 samples when considering the 2–3 biological replicates of each genotype in each harvest.

2.3. Wood chemical analysis of reference samples

This section is described in detail in Supplementary Information Text (SI Text). The 120 selected samples were analyzed for chemical composition following standard analytical methods (wet chemical analysis, HPLC, analytical pyrolysis) to generate reference values used to develop dedicated calibrations to predict wood chemical traits in all the samples ($n = 5799$). Wood chemical traits included: (i) extractives content; (ii) lignin content (Klason lignin, acid-soluble lignin); and (iii) the content of the two most abundant cell wall sugars (glucose, xylose). Analytical pyrolysis was used to assess lignin composition [relative proportion of p-hydroxyphenyl (H), guaiacyl (G) and syringyl (S) units] according to Rodrigues et al. (2001, 1999) and Alves et al. (2006). Except for analytical pyrolysis, at least two technical replicates were performed per sample. For analytical pyrolysis, technical replicates were done only for a few samples to estimate the root mean square error (RMSE) of the method for further comparison with the RMSE of the corresponding NIR calibration.

2.4. Development of NIR calibration models using partial least squares (PLS) regression

R software was used for PLS regression model development (R Core Team, 2015). To perform the calibrations, we used the R package *pls* (Mevik and Wehrens, 2007). Various home-made functions were also used to carry out the calibrations with PLS regression in a cross-validation scheme with an optional detection of potential outlier observations. Moreover, the function “carspls_LOO” was used for automatically selecting a subset of wavenumbers to be included in the PLS regression as proposed by Li et al. (2009). The selection is based on an iterative exclusion of wavenumbers according to their weight in a PLS regression and following an exponential decreasing function. Consequently, the selected wavenumbers are specific to the trait being calibrated. More details about this method are given in Li et al. (2009).

Prior to a final calibration step, we detected potential outlying observations within the 120 reference samples using either box-and-whisker plots or *P-value* thresholds of z-tests on the cross-validation residuals of PLS calibrations. The final calibration step involved splitting of the 120 reference samples into a calibration set ($n = 99$, $\sim 5/6$) and a validation set ($n = 21$, $\sim 1/6$) using Kennard-Stone algorithm (Kennard and Stone, 1969) per harvest on first derivative spectra. Next, outliers detected in the previous step were removed from both calibration and validation data sets. The resulting calibration set was then used to build the model with a leave-one-out (LOO) cross-validation with or without automatic wavenumber selection using the CARS algorithm (Li et al., 2009). The optimal number of components in the PLS regression model was optimized within the cross-validation using Wold's criterion (Li et al., 2002), which was set up at 1. The following statistics were calculated for each model both within the training (cross-validation) and validation sets:

- The coefficient of determination defined as $R^2 = 1 - \left(\frac{RSS}{TSS}\right)$, where RSS is the residual sum of squares (sum of squares of differences between observed and predicted values), and TSS is the total sum of squares (sum of squares of differences between observations and their mean);
- The root mean square error defined as $RMSE = \sqrt{\left(\frac{RSS}{n}\right)}$, where RSS is defined as above and n is the number of observations;

- The ratio of prediction to standard deviation defined as $RPD = \frac{SD}{RMSE}$, where SD is the standard deviation of the observations, and RMSE is defined as above.

The models with best statistics were selected and, when validated, used to predict all samples ($n = 5799$) included in the study.

2.5. Estimation of genetic parameters for NIR-predicted wood chemical traits

The NIR-predicted wood chemical traits were all approximately normally distributed and data transformations were not considered necessary prior to genetic analysis. In order to estimate variance components of traits, linear mixed models (Henderson, 1984) involving spatial effects were fitted using breedR package (Muñoz and Sanchez, 2015) in software R for the analysis of all predicted traits within harvests. Both block and spatial effects account for the environmental variation within the experimental field. Block effects account for global field variations, while spatial effects capture the environmental heterogeneity not accounted by the block effects because of the relatively large size of each block. Furthermore, spectra data have been collected according to the ordered field positions of the trees. So spectra collection date is likely to contribute to the so called spatial variation revealed by the variograms. Accounting for the date effect could help to interpret the spatial effects, if necessary.

For each of the traits, the following mixed model was fitted:

$$y = X\beta + Zu + Rb + Nd + e \quad (1)$$

where y is a vector of individual tree data for a predicted wood chemical trait, β is a vector of fixed effects (over all mean or intercept), u is a vector of random effects of genotypes (genetic effects of genotypes or genotypic values), b is a vector of random effects of blocks, d is a vector of random effects of the dates of NIR spectra collection and e is a vector of residuals. X , Z , R , and N are known incidence matrices relating the observations to the fixed effects in vector β and random effects in vectors u , b , and d , respectively, assuming $u \sim N(0, \sigma_u^2 I)$, $b \sim N(0, \sigma_b^2 I)$, $d \sim N(0, \sigma_d^2 I)$, $e \sim N(0, R)$, where σ_u^2 is the genotypic variance, σ_b^2 is the block variance, σ_d^2 is the date variance, R is the residual covariance matrix, and I is an identity matrix. A spatial residual structure was implemented in order to decompose e into spatially dependent (ξ) and spatially independent (η) residuals (Dutkowski et al., 2002), leading to the following decomposition of R :

$$R = \sigma_\xi^2 [AR1(\rho_{col}) \otimes AR1(\rho_{row})] + \sigma_\eta^2 I \quad (2)$$

where σ_ξ^2 is the spatially dependent residual variance, σ_η^2 is the spatially independent residual variance, I is an identity matrix and $AR1(\rho)$ is a first-order autoregressive correlation matrix.

The mixed model described in model 1 was compared with a model without decomposition of the residual term into spatially dependent and independent effects based on the Akaike information criterion (AIC) and was found to have a lower AIC (i.e., better performance) in all data sets (i.e., ORL2012 and SAV2011 harvests) for all predicted phenotypes. However, spatial trends were not modelled for ORL2010 harvest because the number of genotypes per harvested block was not large enough to capture the within block spatial variation. Moreover, the level of sampling within each block induced heterogeneity in the spatial distribution of the corresponding samples, so estimation of spatial effects over the trial could be biased.

Within each harvest and for each phenotype, reduced models (dropping block or spectra collection date effect) were also fitted and compared to the corresponding full models based on the AIC. Finally, the model yielding the best fit (lowest AIC) was selected for variance component estimation and to adjust the phenotype for non-genetic random effects (block, date, spatially dependent residuals).

Variance components from the selected mixed model were further used to estimate broad-sense heritabilities of the NIR-predicted wood

chemical traits within harvests. Individual tree broad-sense heritability (H_i^2) was calculated using the following equation:

$$H_i^2 = \frac{\sigma_G^2}{(\sigma_G^2 + \sigma_e^2)} \quad (3)$$

where σ_G^2 and σ_e^2 are the genotypic and residual variance components, respectively. Clonal mean broad-sense heritability or clonal repeatability (H_c^2) was calculated as:

$$H_c^2 = \frac{\sigma_G^2}{(\sigma_G^2 + \sigma_e^2/r)} \quad (4)$$

where r is the average number of replicates per genotype for a trait under consideration for a given harvest. Standard errors for heritability were calculated using the Delta method (Lynch and Walsh, 1998). Standard errors were multiplied by 1.96 to construct the 95% confidence interval (CI) for heritability.

Finally, we used genotypes shared between coppice rotations or sites for fair comparisons of genetic parameter estimates and for assessing stability of genetic parameters between rotations as well as between sites or characterizing $G \times E$ interaction. A set of 289 genotypes were shared between rotations within Orleans' trial (ORL2010 vs ORL2012 harvests), while 683 genotypes were shared between sites (ORL2012 vs SAV2011 harvests).

For analysis within harvests based on shared genotypes, the following model was fit:

$$y = X\beta + Zu + e \quad (5)$$

Where y is a vector of individual tree data for a predicted wood chemical trait that was adjusted for block, date and spatially dependent effects with the previously selected mixed model (model 1) and the remaining parameters were assigned as described in model 1.

In order to test and evaluate the extent of $G \times E$ interaction across rotations and sites, the following $G \times E$ mixed model was fitted:

$$y = X\beta + Zu + Mp + e \quad (6)$$

where y is a vector of individual tree data for a predicted wood chemical trait that was adjusted for block, date and spatially dependent effects with model 1, β is a vector of fixed effects (over all mean and rotations or sites), p is a vector of random effects of $G \times E$ interaction. X , Z and M are incidence matrices relating the observations to the fixed effects in vector β and random effects in vectors u and p , respectively, assuming, $p \sim N(0, \sigma_{G \times E}^2 I)$, where $\sigma_{G \times E}^2$ is the $G \times E$ interaction variance. The remaining parameters were assigned as described in model 1. Likelihood ratio tests (LRT) between the full model and a reduced model without $G \times E$ interaction effect were performed to test the significance of $G \times E$ interaction effect. Correlations between adjusted genotype means were also estimated using the Spearman's rank correlation to further characterize the stability of genotype means between rotations or sites. Finally, when the extent of $G \times E$ variance was found to be more than 50% of the genotypic variance (Shelbourne, 1972), a further decomposition of the $G \times E$ interaction was carried out following method 1 of Muir et al. (1992). Such decomposition enables to partition the $G \times E$ sum of squares into scaling effect and genotype rank change.

3. Results

3.1. Variability in wood chemical properties of reference samples

The descriptive statistics for traits analyzed in the laboratory of the 120 reference samples are presented in Table 2. The range of variation in most of the traits analyzed was considerable, and provided the potential to develop reliable calibrations. For example, Klason lignin content ranged from 16.8% to 26.5%, whereas glucose content ranged from 30.6% to 50.3%. Overall, the range of wood chemical traits within our reference data set was 5–21 times the RMSE of the standard

Table 2

Descriptive statistics for lignin monomers (H, G, S), lignin composition (H/G, S/G), lignin content (Klason lignin, Py-lignin, acid-soluble lignin), cell wall sugars (xylose, glucose, xylose/glucose, C5/C6) and extractives analyzed by standard laboratory methods for 120 reference wood samples.

Trait	Unit	RMSE	Min.	Max.	Mean
H-lignin	% Lignin	0.60	2.80	11.0	5.00
G-lignin	% Lignin	0.84	41.20	53.10	46.40
S-lignin	% Lignin	1.01	39.50	54.90	48.60
H/G	fold	0.01	0.05	0.25	0.11
S/G	fold	0.03	0.86	1.34	1.06
Klason lignin	% CWR	1.61	16.80	26.5	21.50
Py-lignin	% CWR	1.49	20.20	27.00	23.10
Acid-soluble lignin	% CWR	0.32	4.60	7.00	6.10
Xylose	% CWR	1.15	13.10	18.70	15.30
Glucose	% CWR	1.87	30.60	50.30	40.40
Xylose/Glucose	fold	0.03	0.29	0.48	0.38
C5/C6	%	1.14	17.9	29.30	23.90
Extractives	% DW	0.54	6.20	17.70	10.40

DW: dry weight; CWR: cell wall residue (extractives-free dry weight); RMSE: root mean square error of the standard methods for replicate analysis.

methods, making this reference data set acceptable for building near-infrared multivariate calibration models (Table 2).

3.2. Calibration, validation and prediction

The absorption spectra modalities (with or without pretreatment) and reference values of the reference samples were used to develop NIR calibration models at a global scale for a majority of the traits, except for lignin contents, where site specific models showed higher predictive performance than the global ones (Table 3). The reference values in the calibration and validation data sets for the wood chemical traits of black poplar were comparable (i.e., had similar means and ranges), which means that reliable models can be developed and effectively verified (Fig. S3).

Summary statistics that demonstrate the performance of the models in calibration and validation data sets are reported in Table 3 and plots of the predicted versus measured component values for selected calibrations are shown in Fig. S4. Pretreated spectral data provided better calibrations than raw spectra. Automatic wavenumber selection improved model performance, for some of the traits, compared with full range.

Global calibration models developed for the prediction of H-lignin, lignin H/G and S/G ratios, xylose/glucose, C5/C6 and extractives were good, with coefficients of determination (R^2) ranging from 0.75–0.91 and 0.72–0.83 in calibration and validation data sets, respectively (Table 3, Fig. S4). The R^2 values for calibration and validation sets of glucose were 0.76 and 0.64, respectively. The models for G-lignin and S-lignin had moderate performance in cross-validation ($R_{cv}^2 = 0.68$ and 0.64, respectively), while the model for S-lignin showed a higher accuracy of prediction ($R_{val}^2 = 0.77$). The model for xylose showed inadequate fit in the calibration data set ($R_{cv}^2 = 0.48$) as well as a poor prediction performance in the validation data set ($R_{val}^2 = 0.29$).

On the other hand, global models for lignin content (Klason lignin, Py-lignin and acid-soluble lignin) were very specific (i.e., they were good in predicting samples included in the models but very poor in predicting samples of an independent validation set) (Table 3). Therefore, we followed the site specific approach to model these characteristics (Table 3, Fig. S4). For Klason lignin, the model developed for Orleans site ($R_{cv}^2 = 0.78$, $R_{val}^2 = 0.60$) had a better fit, whereas good models for Py-lignin and acid-soluble lignin were obtained at Savigliano ($R_{cv}^2 = 0.79$, $R_{val}^2 = 0.73$ and $R_{cv}^2 = 0.77$, $R_{val}^2 = 0.79$, respectively).

With the exception of Klason lignin, Py-lignin, acid-soluble lignin and xylose global models, the remaining global models described in Table 3 and Fig. S4 were used to predict wood chemical properties for

Table 3

NIR calibration models (leave-one-out cross-validation) and validation statistics for wood chemical properties evaluated on 120 reference samples. For trait abbreviations see the caption of Table 2.

Trait	Model type	Calibration set (n = ~5/6)								Validation set (n = ~1/6)				
		nlambda	Pretreatment	nbcomp	R _{cv} ²	RMSE _{cv}	RPD _{cv}	nobs	nb. outliers	R _{val} ²	RMSE _{val}	RPD _{val}	nobs	nb. outliers
H-lignin	Global	Full range	der2	9	0.75	0.80	2.0	91	8	0.80	0.89	2.3	21	0
G-lignin	Global	29	der2	5	0.68	1.26	1.8	94	5	0.51	1.33	1.5	20	1
S-lignin	Global	Full range	norm-der2	13	0.64	1.25	1.7	90	9	0.77	1.02	2.2	20	1
H/G	Global	652	der2	8	0.82	0.02	2.4	92	7	0.83	0.02	2.5	19	2
S/G	Global	947	norm-der2	12	0.84	0.03	2.5	91	8	0.72	0.04	2.0	21	0
Klason lignin	Global	Full range	der1	5	0.61	1.17	1.6	91	8	0.27	1.44	1.2	20	1
	Site: ORL	Full range	dt	6	0.78	0.94	2.2	56	10	0.60	1.31	1.6	13	1
Py-lignin	Site: SAV	Full range	der2	5	0.25	1.43	1.2	33	0	-4.33	1.22	0.5	7	0
	Global	Full range	norm-der2	7	0.75	0.59	2.0	97	2	-0.18	0.78	1.0	21	0
	Site: ORL	Full range	norm-der2	7	0.72	0.65	1.9	65	1	-1.43	0.87	0.7	14	0
Acid-soluble lignin	Site: SAV	Full range	der1	8	0.79	0.39	2.2	26	7	0.73	0.35	2.1	6	1
	Global	Full range	der2	7	0.61	0.23	1.6	92	7	0.35	0.29	1.3	18	3
	Site: ORL	Full range	norm-der2	6	0.49	0.25	1.4	61	5	0.21	0.29	1.2	13	1
Xylose	Site: SAV	Full range	norm-der2	8	0.77	0.16	2.1	30	3	0.79	0.15	2.4	6	1
	Global	Full range	der1	6	0.48	0.73	1.4	88	11	0.29	0.85	1.2	21	0
Glucose	Global	46	norm-der1	6	0.76	1.49	2.0	94	5	0.64	1.25	1.7	18	3
Xylose/Glucose	Global	28	norm	8	0.79	0.02	2.2	90	9	0.75	0.02	2.0	20	1
C5/C6	Global	129	der2	6	0.85	0.89	2.6	91	8	0.81	1.08	2.4	21	0
Extractives	Global	326	der1	9	0.91	0.74	3.3	91	8	0.74	1.03	2.0	20	1

nlambda: number of selected wavenumbers; nbcomp: number of PLS components; R_{cv}²: coefficient of determination of cross-validation; RMSE_{cv}: root mean square error of cross-validation; RPD_{cv}: ratio of performance to deviation of cross-validation; nobs: number of samples statistically analyzed; R_{val}²: coefficient of determination of validation; RMSE_{val}: root mean square error of validation; RPD_{val}: ratio of performance to deviation of validation; norm: normalized spectra; dt: detrending spectra; der1: first derivative spectra; der2: second derivative spectra; norm-der1: first derivative on normalized spectra; norm-der2: second derivative on normalized spectra; Full range spectrum: 8000–4000 cm⁻¹.

the entire sample set (n = 5799) to study phenotypic variability, degree of genetic control and G × E interaction. Klason lignin at Orleans and Py-lignin and acid-soluble lignin at Savigliano were predicted with the site specific models because of the poor prediction performance of their corresponding global models. On the other hand, Klason lignin at Savigliano and Py-lignin and acid-soluble lignin at Orleans were not predicted since their corresponding site specific models had poor performances, especially in validation sets. Xylose was not predicted since the quality of the model was considered poor as mentioned before.

Boxplots of the distributions of NIR-predicted wood chemical traits (without adjustment for micro-environmental effects) are presented in Fig. S5. The range of phenotypic variation in most predicted wood traits was considerable. For example, the predicted Klason lignin content ranged from 16.1% to 27.9%, whereas predicted glucose content ranged between 30.2% and 49.7%. All the predicted values were in line with the results obtained for the reference data set (i.e., they were pretty much close to the limits or within the range of variation observed for the reference data set) (Table 2). Moreover, based on comparisons of the RMSE of the models (Table 3) to the RMSE of the standard methods (Table 2), the uncertainties associated with the predictions can be regarded as acceptable. It is worth mentioning, however, that the predicted S/G values (0.69–1.43) didn't fall within the range of values reported for other populations of *P. nigra* (1.3–2.1) (Guerra et al., 2013) and *P. trichocarpa* (1.5–2.4) (Guerra et al., 2016) despite variations of almost the same magnitude.

3.3. Variance components and broad-sense heritability of wood chemical traits within harvests

Due to some unbalance in genotype representation across harvests, the genetic analysis of individual harvests using only the shared genotypes was done to ensure fair comparisons of genetic parameters for the same traits. Thus, a set of 289 genotypes were shared between ORL2010 and ORL2012 harvests (i.e., between coppice rotations), while 683 genotypes were shared between ORL2012 and SAV2011 harvests (i.e., between sites).

Based on analysis using 289 genotypes, high clonal repeatability

(H_c^2) values were found for lignin monomers (H, G, S) (0.74 ± 0.05 to 0.81 ± 0.04), lignin composition (H/G, S/G) (0.75 ± 0.05 to 0.81 ± 0.04), Klason lignin (0.75 ± 0.05 to 0.80 ± 0.04) and cell wall sugars (0.72 ± 0.06 to 0.80 ± 0.04) in the 2 rotations (Fig. 1, Table S1). The exception was extractives content, for which the H_c^2 values were moderate to high (0.57 ± 0.09 to 0.72 ± 0.06). Using 683 genotypes, high H_c^2 values were found for all traits except extractives, namely, lignin monomers (0.74 ± 0.04 to 0.88 ± 0.02), lignin composition (0.77 ± 0.03 to 0.89 ± 0.01) and cell wall sugars (0.70 ± 0.04 to 0.81 ± 0.02) in the two sites (Fig. 2, Table S2). For extractives, the H_c^2 values were moderately high (0.62 ± 0.05 to 0.70 ± 0.04). At Savigliano, where trees grew more rapidly, the genetic control over all the chemical traits was generally stronger with the exception of C5/C6. For example, H_c^2 for lignin S/G ratio was higher (0.89 ± 0.01) at Savigliano compared to $H_c^2 = 0.77 \pm 0.03$ at Orleans. These differences in heritability of the same traits between the two sites can be explained by scale effects (i.e., both increased expression of genetic variation and decreased residual variation at Savigliano as compared to Orleans) (Table S2). In comparison to site, rotation effects on H_c^2 were rather low for all traits with the exception of extractives, suggesting differences in magnitude of G × E interaction between rotations and sites. For extractives content, H_c^2 varied more between rotations than between sites. Next, in order to reflect the genetic variation in wood traits that could be present in the entire population, we repeated the genetic analysis using all genotypes available in each harvest.

To estimate genetic parameters for the NIR-predicted wood chemical traits for each single harvest using all genotypes available, data were analyzed with a full mixed model accounting for spatial effect.

Overall, the clonal repeatability estimates were highly comparable between the two analyses, i.e., all genotypes versus shared genotypes (Figs. 1–2, Table S1 and S3). We conclude that the genotypes shared between the harvests were adequate enough to capture the genetic variation existing in the entire population. This is interesting because we would not miss important information when analysing the G × E interaction both across rotations and sites using the shared genotypes.

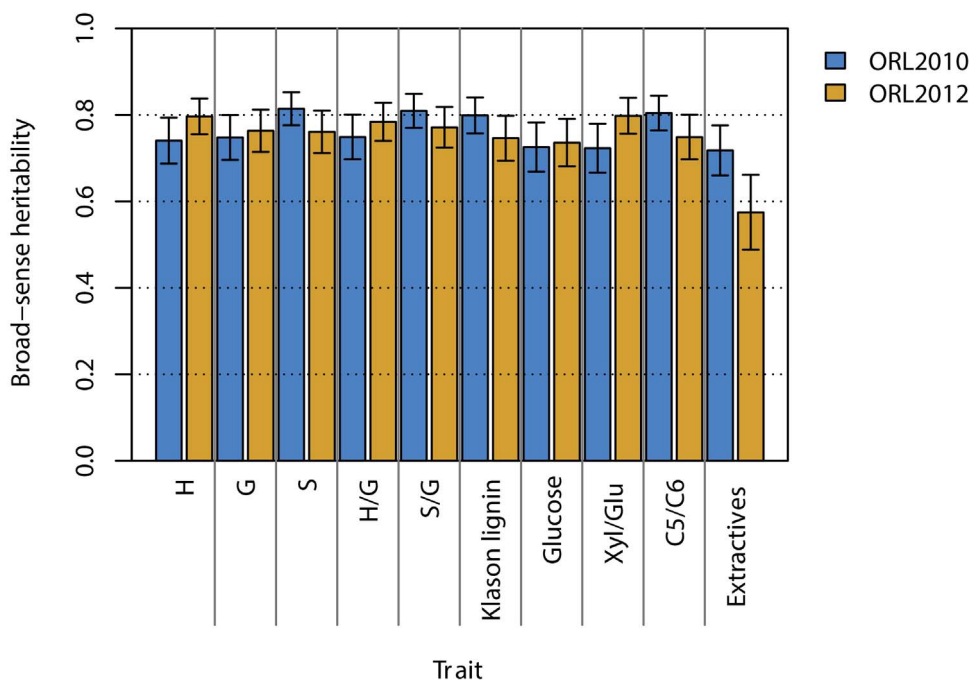


Fig. 1. Estimated clonal mean broad-sense heritability (clonal repeatability) (H_c^2) with error bars corresponding to the 95% confidence intervals for NIR-predicted wood chemical traits evaluated over two successive 2-yr rotations (ORL2010 and ORL2012) in a clonal trial at Orleans (France). For trait abbreviations see the caption of Table 2.

3.4. Genotype × environment ($G \times E$) interaction effect on wood chemical traits

To assess the stability of genetic parameters or characterize $G \times E$ interaction for the NIR-predicted wood chemical traits, genotypes shared between rotations (ORL2010 vs ORL2012 harvests) or between sites (ORL2012 vs SAV2011 harvests) were used. Two strategies were adopted to assess $G \times E$ interaction including estimation of variance components and correlation between environments.

Combined mixed model analysis of variance of 289 genotypes

evaluated across rotations at Orleans showed that the $G \times E$ interaction effect was significant (LRT P -values < 0.001, 0.01) for all traits, except for H-lignin and lignin H/G ratio (Table S4). However, the magnitude of the $G \times E$ interaction variance was rather low, compared to the genotypic variance component for all traits. The $G \times E$ interaction variance component explained only 1–18% of the total phenotypic variance, whereas the genotype main effect accounted for 30–53% (Fig. 3, Table S4). Based on the 683 genotypes tested across sites, for all traits highly significant (LRT P -value < 0.001) $G \times E$ interaction was found and the GxE variance reached more than 50% of the correspond-

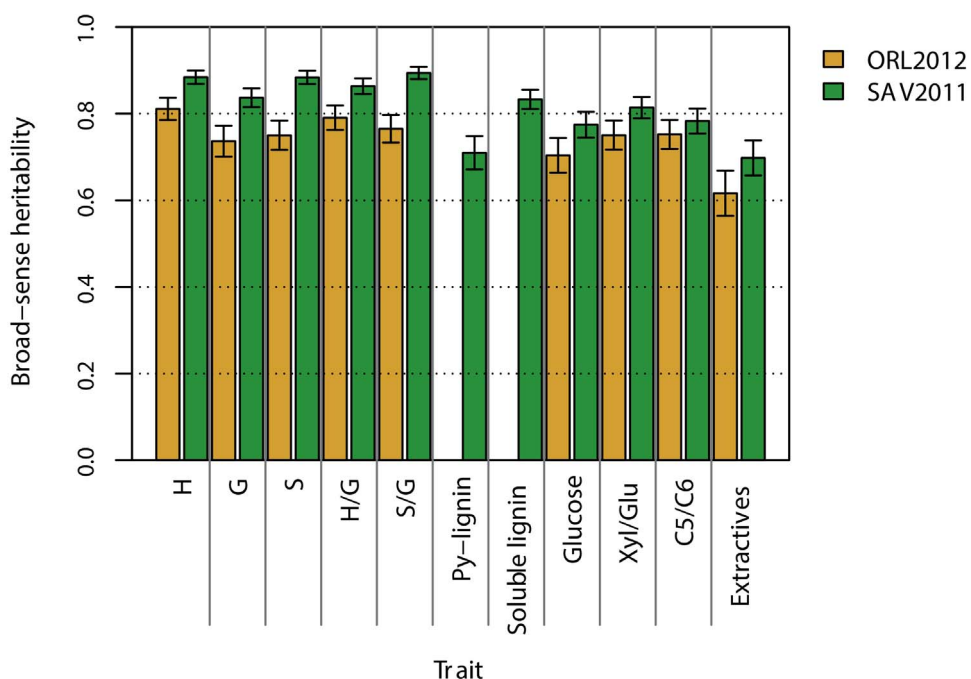


Fig. 2. Estimated clonal mean broad-sense heritability (clonal repeatability) (H_c^2) with error bars corresponding to the 95% confidence intervals for NIR-predicted wood chemical traits evaluated at two contrasting sites (Orleans, France: ORL2012; Savigliano, Italy: SAV2011). Trees were grown over two successive 2-yr rotations at Orleans (2008–2009, 2010–2011) and 1-yr and 2-yr rotations at Savigliano (2008, 2009–2010). Results are based on the data from the second rotations at the two sites. For trait abbreviations see the caption of Table 2.

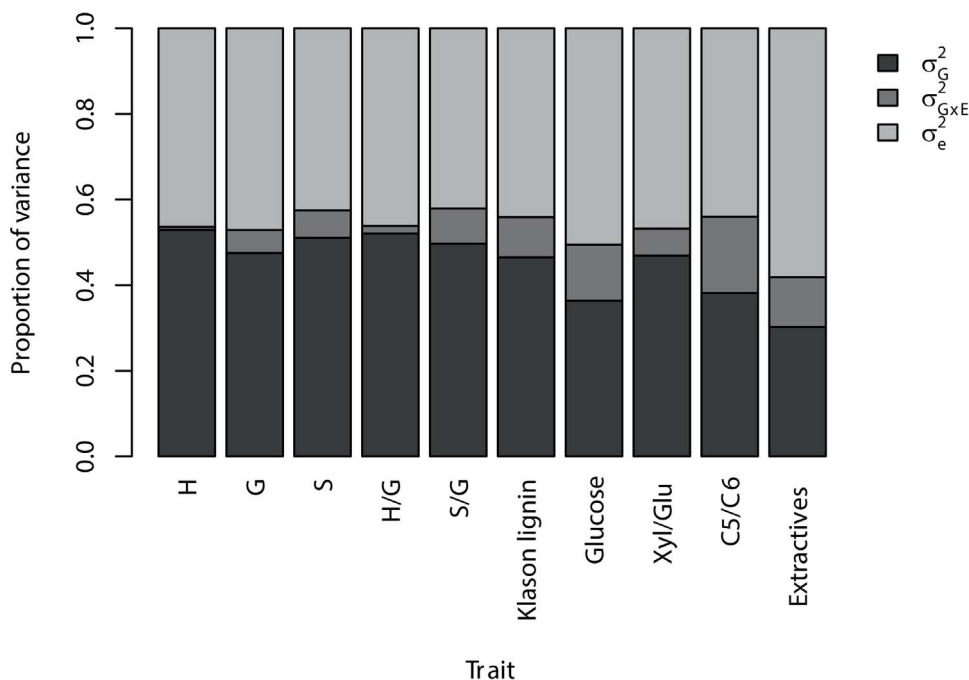


Fig. 3. Decomposition of total phenotypic variance for NIR-predicted wood chemical traits evaluated over two successive 2-yr rotations (ORL2010 and ORL2012) in a clonal trial at Orleans (France). Stacked barplot of the percentage of the total phenotypic variance explained by the genotype main effect (σ_G^2), genotype \times environment (G \times E) interaction effect ($\sigma_{G \times E}^2$) and residual effect (σ_e^2) variance components for 10 wood chemical traits and using 289 shared genotypes. For trait abbreviations see the caption of Table 2.

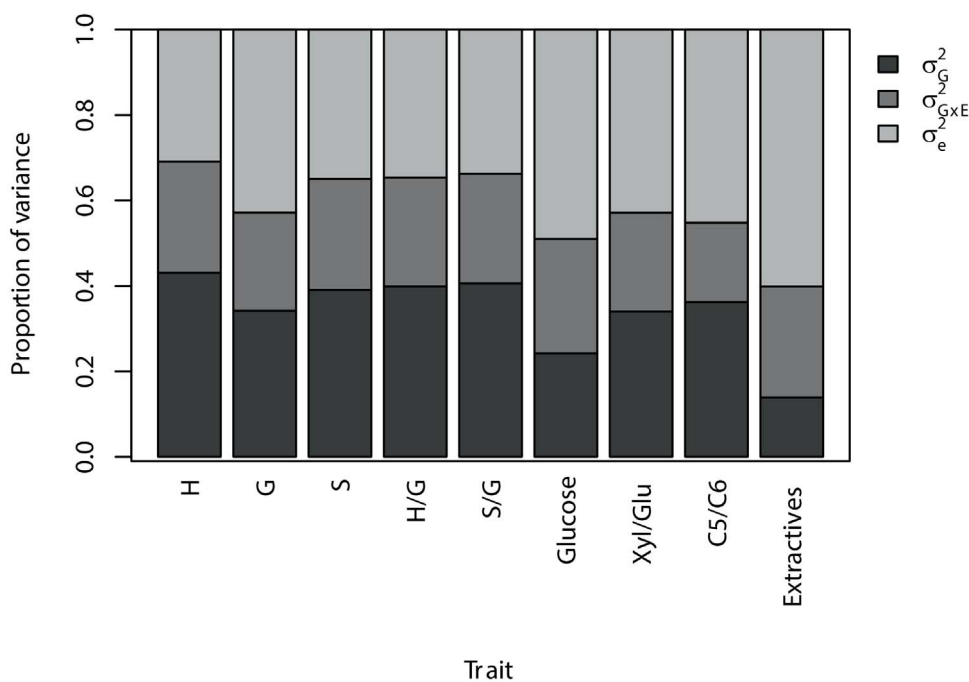


Fig. 4. Decomposition of total phenotypic variance for NIR-predicted wood chemical traits evaluated at two contrasting sites (ORL2012: Orleans, France; SAV2011: Savigliano, Italy). Stacked barplot of the percentage of the total phenotypic variance explained by the genotype main effect (σ_G^2), genotype \times environment (G \times E) interaction effect ($\sigma_{G \times E}^2$) and residual effect (σ_e^2) variance components for 9 wood chemical traits and using 683 shared genotypes. For trait abbreviations see the caption of Table 2.

ing genetic variance (Fig. 4, Table S5). Importantly, for glucose and extractives, the G \times E interaction variance component (27% and 26%, respectively) was even larger than the genetic variance component (24% and 14%, respectively) and this was somewhat mirrored in the relative values of the clonal repeatability, especially for extractives. Compared to extractives, glucose content is a key wood chemical trait in term of bioethanol production. During bioethanol production, glucose is released from cellulose in the plant cell walls via enzymatic hydrolysis of lignocellulosic biomass and could then be converted into

bioethanol via fermentation. To assess if the observed G \times E interaction for glucose in particular and all biochemical traits in general would have practical implications for poplar tree breeding for bioethanol production, it is noteworthy to further decompose the corresponding interaction variance components. Because the G \times E interaction was in general more evident over sites than over rotations for all traits, we sought to zoom into the nature of the G \times E interaction across sites.

Thus, G \times E interaction was dissected according to the method 1 described by Muir et al. (1992). Results of the partitioning of the G \times E

Table 4

Partitioning of genotype \times environment (G \times E) interaction sum of squares (SS G \times E) for NIR-predicted wood chemical traits evaluated at two contrasting sites and using 683 shared genotypes according to Method 1 of Muir et al. (1992). Proportion of genotypes explaining 50% of the SS G \times E was calculated according to their relative ecovalence (Lin et al., 1986). For trait abbreviations see the caption of Table 2.

Trait	$\sigma_{G\ ORL}$	$\sigma_{G\ SAV}$	$\frac{\sigma_G^2}{\sigma_{G \times E}^2}$	r_s	% SS G \times E		Proportion of genotypes explaining 50% of G \times E SS
					Scale effect	Re-ranking	
H-lignin	0.52	0.79	1.66	0.56	14.61	85.39	12.30
G-lignin	0.80	0.77	1.49	0.46	1.46	98.54	10.25
S-lignin	1.29	1.83	1.50	0.51	6.53	93.47	10.69
H/G	0.01	0.02	1.57	0.53	11.33	88.67	13.47
S/G	0.05	0.07	1.59	0.54	2.52	97.48	9.22
Glucose	1.24	1.26	0.91	0.34	0.23	99.77	10.25
Xylose/Glucose	0.01	0.01	1.47	0.45	0.02	99.98	11.71
C5/C6	0.76	0.77	1.95	0.49	0.01	99.99	11.13
Extractives	1.06	0.68	0.54	0.23	14.92	85.08	10.69

r_s : Spearman's rank correlation coefficient.

interaction sum of squares into sources due to scaling effects (heterogeneity of variances) and re-ranking indicated that the G \times E interaction for all traits was dominated by changes in genotype ranking over the two sites (Table 4). Nevertheless, it appeared that only 9–13% of the genotypes, which correspond to the most interactive ones, were found to explain 50% of the G \times E interaction sum of squares. Whereas the impact of G \times E interaction seemed higher for glucose and extractives on the basis of the relative magnitude of their variance components, the proportion of interactive genotypes were found to be quite similar to those for other traits, suggesting less practical importance of the observed interaction. Extractives had some level of scale effect (14.9%) and still high re-ranking (85.1%), whereas glucose had little or no scale effect (0.23%) and high re-ranking (99.77%) (Table 4). Similarly, the genetic variances, expressed in terms of genetic standard deviation (σ_G), were not similar between the two sites for extractives, with higher variation at Orleans (i.e., some level of scale effect), whereas, for glucose, the genetic variances were more homogeneous between Orleans and Savigliano (1.24 and 1.26, respectively) (i.e., little or no scale effect). Although H-lignin and extractives had similar patterns of partitioning of G \times E interaction sum of squares, the Spearman's rank correlation was much weaker for extractives, which was consistent with the relatively stronger G \times E effect on this particular trait, as shown by the ratio of σ_G^2 to $\sigma_{G \times E}^2$ (Table 4).

Furthermore, we assessed the stability of genotype ranking across rotations or sites on a genotypic mean basis for each wood trait using Spearman's rank correlation coefficients (r_s). Thus, (r_s) between the two rotations were stronger than 0.60 ($r_s = 0.64 - 0.71$) for all traits except C5/C6, glucose and extractives (Table S4), which was consistent with the relatively low level of G \times E interaction observed across rotations for most traits (1–9%). For C5/C6, glucose and extractives, the correlations were in the range of 0.48–0.50, which was consistent with the relatively higher proportion of G \times E interaction variances for these three traits (12–18%) (Fig. 3, Table S4). By contrast, the Spearman's rank correlations of genotypic means between the two sites were lower than 0.60 for most of the traits ($r_s = 0.45 - 0.56$), and has turned out to be much weaker for extractives ($r_s = 0.23$) and glucose ($r_s = 0.34$) (Table 4), which corroborated the relatively high level of G \times E interaction observed across sites compared to across rotations (Fig. 4, Table S5).

4. Discussion

A study of this scale would not have been possible using the standard method of wood compositional analysis because of the high cost and time required. For example, to analyze the 120 reference samples in two technical replicates using the wet chemistry method, it took about two months. It means that, it would have taken around 8 years to analyze about the 6000 samples included in our study. A way

to circumvent such technical limitation is to use an attractive technique that combines NIR spectroscopy with multivariate statistical analysis. NIR spectroscopy is an inexpensive and high-throughput technique for phenotyping large-scale wood samples required for the genetic analysis of biofuel related traits and, consequently, it can provide the opportunity to select or develop biofuel-type poplar clones. Nevertheless, NIR spectroscopy is an indirect method which is reliable only if calibration models are provided. In this study, chemical composition data from standard methods and NIR spectra of reference samples were used to develop and validate calibrations taking into consideration the phenotypic variation induced by multi-environment evaluation. The models based on a 120-samples reference set were then used to predict the composition of the 5799 black poplar samples covering the range of the species in Western Europe. Using NIR predictions, we evaluated their genetic variability and the extent of G \times E interaction across coppice rotations and sites. To our knowledge, this is the first work to evaluate large-scale clonal trials of *P. nigra* for wood chemical traits using an indirect method of measurement.

4.1. Calibration reliability

When global calibration models developed for the prediction of 6 wood chemical traits (H-lignin, S/G, H/G, xylose/glucose, C5/C6, extractives) in samples of European black poplar were tested on an independent validation data set, they gave good fits, suggesting their potential use in genetic analysis of large data sets or for ranking of genotypes with respect to their predicted phenotypic performances in initial selection steps in breeding programs. RPD values of ~ 1.5 indicate that the models are acceptable as initial screening tools, whereas RPD greater than 2.5 suggest that the models are good for screening candidates in breeding programs (Yeh et al., 2005). Although site effects were apparent on some traits such as xylose/glucose ratio, we were able to develop non-site specific calibrations for such parameters.

By contrast, global models for lignin content (Klason lignin, Py-lignin, acid-soluble lignin) showed clearly poorer performance in the validation set than in the calibration set. There is no obvious explanation for these apparent differences in global model performance between calibration and validation sets. We further examined if model fit was better for site specific calibration than global ones for lignin content and found that the Klason lignin model had a good fit at Orleans, while Py-lignin and acid-soluble lignin models had good fits at Savigliano, which does explain why we could not have global models for these characteristics.

Only a few studies have investigated the efficiency of NIR calibration models for prediction of poplar wood composition and these focused on hybrid poplars instead of natural populations. Robinson and Mansfield (2009) used NIR spectra of 267 wild and transgenic

hybrid poplar samples coupled with a modified thioacidolysis protocol for predicting lignin monomer proportions (S, G, and H). The authors reported highly accurate calibrations with prediction R^2 values of 0.96, 0.96, and 0.71 for S, G and H, respectively. More recently, Zhou et al. (2011) used Fourier transform infrared spectroscopy (FTIR) and acetyl bromide method to develop a calibration model for predicting lignin content in hybrid poplar wood samples. They reported a strong calibration with cross-validation R^2 of 0.81 and prediction R^2 of 0.88. Our global model for H-lignin had higher prediction R^2 than the local model developed by Robinson and Mansfield (2009). However, we found lower prediction R^2 values for the predominant G and S lignin monomers. Compared with the lignin model developed by Zhou et al. (2011), our local models for Klason lignin, Py-lignin and acid-soluble lignin had slightly lower R^2 values. The differences between these previous studies and our work may be largely related to differences in study population and standard laboratory methods. However, in this study spectra were recorded on non-debarked and non-extracted wood samples because it is practically difficult to debark and extract a large amount of samples ($n = \sim 6000$). Although the presence of bark and extractives may disturb the spectra, we still attained sufficiently accurate models for a majority of the traits analyzed. Furthermore, the R^2 value may be misleading as it depends not only on the model error but also on the range of variation of the trait of interest within or across sites. For example, in this study the effect of site on the range of variation of some of the traits analyzed was quite high (Fig. S5). Consequently, different R^2 values can be obtained for calibrations for the same trait at the two sites while still having the same prediction error. Higher R^2 values can be obtained for calibrations at site that is more variable than other.

4.2. Variabilities, $G \times E$ interactions and broad-sense heritability of wood chemical properties

In this study, using NIR predictions of a large number of wood samples from *P. nigra* clonal trials, we assessed variabilities, $G \times E$ interaction and broad-sense heritability for wood chemical traits. The range of phenotypic variation in most NIR-predicted wood chemical traits in the black poplar populations studied was substantial. Guerra et al. (2013) used Pyrolysis molecular beam mass spectrometry (pyMBMS) to determine C6 sugars, total lignin content and S/G ratio in wood samples of 2-yr-old trees, representing 17 open-pollinated families of *P. nigra*. Porth et al. (2013) used wet laboratory approaches to determine xylose, glucose, Klason lignin and acid soluble lignin in wood samples of 9-yr-old trees, representing natural populations of *P. trichocarpa*. The range of variation observed for predicted glucose content (30.2–49.7%) in the present study is in accordance with that reported for C6 sugars (27.7–39.7%) by Guerra et al. (2013) and for glucose (40.7–61.7%) by Porth et al. (2013). We obtained predicted Klason lignin content (16.1–27.9%) that is well comparable to the results of total lignin content (Klason and soluble lignin) reported by Guerra et al. (2013) (19.5–26.5%) and Porth et al. (2013) (14.7–25.7%). The range of variation of the predicted lignin S/G ratio (0.69–1.43) described in this study doesn't mirror the range between 1.3 and 2.1 reported by Guerra et al. (2013) despite almost the same magnitude of variations, which might arise from the differences in the standard methods of lignin monomers determination.

The effects of harvest on some of the predicted wood chemical traits are evident in Fig. S5. This motivated us to ask whether there is a significant influence of $G \times E$ interaction on the wood chemical traits. Understanding the magnitude and nature of $G \times E$ interaction would be useful for establishing breeding objectives. To estimate the importance of $G \times E$ interaction, we examined variance contributions of $G \times E$ interaction for the wood traits and correlations of same traits between environments based on genotype means. In this study, significant $G \times E$ interaction was observed across rotations as well as across the two sites for a majority of the traits assessed, suggesting differential responses of

genotypes to the environmental conditions. The $G \times E$ interaction variance component accounted for a lower proportion of the total variance across rotations than across sites and this was consistent with the rank correlations of genotype means between rotations and sites obtained for most traits examined. Together, it implies that genotype ranking was relatively more maintained between rotations than between sites. The observed differences in the magnitude of interactions between rotations and sites were not surprising, since the clonal trials were established at two contrasting sites, particularly in terms of soil fertility. Savigliano is characterized by a higher soil fertility compared to Orleans (Guet et al., 2015). Given the differences in edaphic factors between the trial sites, the significant $G \times E$ effect revealed for wood chemical traits across sites could result indirectly from the effects of edaphic factors on tree growth.

When the contributions of $G \times E$ interaction and genotype main effects to the total phenotypic variances of predicted wood traits were compared, all the traits had a higher percentage of variance due to genetic variance component, suggesting less consequences of interaction in poplar tree breeding for improved wood quality. The exceptions were the glucose and extractives contents across sites, for which the $G \times E$ variance components were larger than the genetic variance components, which was also consistent with the relatively lower rank correlations of genotype means between sites for these two traits. Nevertheless, the partitioning of the $G \times E$ sum of squares revealed that the $G \times E$ effect was mainly caused by a few interactive genotypes as for the other traits. This suggests that the interaction would have less consequences in the poplar tree breeding programs for biofuel production because there exists a high possibility to identify genotypes with stable wood quality across the two sites. To test this assumption, we have further computed the relative loss in genetic gain that would arise when selecting the best 5% genotypes for some relevant traits (S/G, H/G, glucose, xylose/glucose, C5/C6) on their genotype mean across the 2 sites instead of their genotype mean within each targeted site. We found that this loss would be fairly low relatively to the maximum expected gain in the two targeted sites (13.1 and 14.3% on average at Orleans and Savigliano, respectively).

To date, only a few studies have investigated the effect of $G \times E$ interaction on wood chemical properties, especially in poplars. Kačík et al. (2012) studied poplar hybrid clones and reported the presence of significant clone \times site interaction for wood chemical traits (lignin content, cellulose, holocellulose, extractives, S/G ratio). However, the authors did not provide further information about the implications of the observed interaction for poplar tree breeding for wood quality. Similarly, Zhang et al. (2015) found significant clone \times site interaction for lignin content and extractives in triploid hybrid clones of *P. tomentosa*. However, clone by site variance exceeded clonal variance only for holocellulose content, for which the authors did not detect significant interaction, and not for lignin content or extractives.

Consistent with the observed $G \times E$ interaction, extractives content showed relatively low within-site broad-sense heritability estimates in this study. Compared to the main wood components, extractives content may be of less interest as a direct selection trait in poplar breeding programs for biofuel production. Since chemical analysis was carried out on non-debarked wood samples in the present study, we wondered if such particular pattern of variation for extractives content would be somehow related to variation in bark proportion. To test this hypothesis, we sought to use the diameter of the samples as a proxy of bark proportion: samples with relatively large diameter are expected to have less bark, and consequently, less extractives. Clearly, extractives content tended to decrease with increasing tree diameter (not shown). We thus extended our $G \times E$ analyses to tree circumference at 1 m aboveground at harvest in order to check if it could explain the particular pattern of variation observed for extractives in comparison with the other wood chemical traits. Interestingly, we found that, albeit highly significant, the $G \times E$ interaction effect accounted for much less variation than the genotype main effect, resulting in G to $G \times E$

variances ratio of 3.90 and 1.31, as well as rank correlations between genotype means of 0.68 and 0.53 across rotations and sites, respectively (Table S6). This pattern of $G \times E$ across rotations and sites was pretty much consistent with the pattern observed for all wood chemical traits, but did not explain the exceptionally interactive aspect of extractives. We thus conclude that, of all the traits evaluated in this study, extractives content was the most interactive trait with moderate heritability and we found no evidence for our hypothesis that the $G \times E$ effect on extractives is confounded by $G \times E$ effect on tree circumference. This result is also supported by the fact that we developed a good global calibration for extractives regardless of the differences in sample bark content between the two sites.

We also quantified the extent of genetic variation present within the European populations of black poplar in the clonal trials using the NIR predictions. Broad-sense heritability was estimated at both individual tree and clonal mean levels. For clonal selection, clonal mean broad-sense heritability (clonal repeatability) is more meaningful. Genetic analysis with NIR predictions revealed that the studied wood chemical traits were under moderate to high genetic control. However, care must be taken when interpreting the heritability estimates reported in the present study because they were estimated from phenotypic data that had been adjusted for within-site non-genetic random effects like block, date and spatially dependent residuals. Consequently, they were over-estimated to an extent that corresponds to an omission of non-genetic random variances in the denominator of the heritability ratio when estimated from the first model (using all genotypes within each harvest, as reported in Table S3). Still, our results suggested that satisfactory genetic gains could be realized in wood chemical traits through clonal selection using a fairly low number of replicates (2.7–2.8 per genotype on average) when NIR analysis is integrated in a breeding program to evaluate large sets of candidate clones. In this regard, the information produced in this research could be used for screening individuals with desirable traits from large-scale clonal trials as future potential parent trees for hybrid breeding programs aimed at cellulosic ethanol production. A general trend was observed for the studied traits in terms of clonal repeatability. Lignin monomers and lignin composition had the highest values, followed by lignin contents, cell wall sugars and extractives (Figs. 1 and 2, Tables S1 and S2). However, the estimated clonal repeatability differed more between sites than between rotations for the same traits, which was in agreement with the $G \times E$ interaction results. The higher clonal repeatability estimates obtained for most of the traits at Savigliano may be explained by the existence of a relatively favourable growth conditions for poplar trees at this site, which resulted in both increased expression of genetic variation and reduced residual variation. Savigliano could be a suitable growth site to apply the clonal evaluation as it provided the genotypes relatively suitable conditions for expressing their genetic potential compared to Orleans.

Using direct method of measurements, previous studies in *P. nigra* (Guerra et al., 2013) and *P. trichocarpa* (Guerra et al., 2016; Porth et al., 2013; Wegrzyn et al., 2010) have also shown that wood chemical properties are under moderate to high genetic control. For example, Guerra et al. (2013) studied 17 cloned open-pollinated families of *P. nigra* and reported individual broad-sense heritability (H_i^2) values of 0.46, 0.58 and 0.70 for C6 sugars, lignin and S/G, respectively. In the current report, the estimated H_i^2 values of 0.47 ± 0.05 – 0.55 ± 0.04 , 0.52 ± 0.07 – 0.59 ± 0.06 and 0.55 ± 0.04 – 0.75 ± 0.03 for glucose, Klason lignin and S/G, respectively, compares favourably well with H_i^2 values reported by these previous authors (Figs. 1 and 2, Tables S1 and S2). More recently, Guerra et al. (2016) studied *P. trichocarpa* clones sampled in provenances and reported the clonal repeatability (H_c^2) estimates of 0.22, 0.33 and 0.81 for C6 sugars, lignin and S/G, respectively, with an average number of 3 biological replicates per clone. In comparison with the results of S/G reported by these authors, we found similar H_c^2 values for S/G (0.77 ± 0.03 – 0.89 ± 0.01) (Figs. 1 and 2, Table S1 and S2). Porth et al. (2013) studied the narrow-sense heritability of several wood properties in natural popula-

tions of *P. trichocarpa* using molecular markers to measure relatedness and reported values of 0.46, 0.66, 0.97 for glucose, Klason lignin and soluble lignin, respectively. We found higher clonal repeatability for glucose (0.70 ± 0.04 – 0.77 ± 0.03) and Klason lignin (0.75 ± 0.05 – 0.80 ± 0.04), but a lower value for acid-soluble lignin (0.83 ± 0.02), indicating that acid-soluble lignin may be under relatively lower genetic control in *P. nigra* than *P. trichocarpa* (Figs. 1 and 2, Table S1 and 2).

4.3. Adapting the NIR method to clonal trials

An initial step to harness the standing genetic variation in poplar is to evaluate natural populations in multi-site clonal trials. This allows to study the relative importance of genetic, environment and $G \times E$ interaction on important biomass production and biomass composition related traits. In parallel, screening good candidates from clonal trials as future parents would increase the genetic diversity available for breeding poplar trees for cellulosic ethanol production. The goal of bioenergy poplar breeding program is to simultaneously improve biomass production and biomass composition. To incorporate wood quality traits into breeding programs, however, tree breeders need low-cost and high-throughput techniques for determination of biomass composition. Standard methods for analysis of biomass composition such as wet chemistry are useful for evaluating small sample sets, but they have limitations to be used in tree breeding programs, where screening of a large number of samples is mandatory to identify those possessing desirable traits. Standard methods are laborious, costly and time consuming. An alternative way is to use NIR spectroscopy coupled with multivariate statistical approaches. NIR spectroscopy is a high-throughput technique for screening a large population. It is easy to operate, allows non-destructive analysis, needs little sample preparation, provides reliable information, requires less time and minimal cost for assessing large number of samples and captures multiple features of the samples with one operation (Lupoi et al., 2014).

The moderate to high heritability estimates and the detection of $G \times E$ interaction in this study are encouraging for NIR determination of wood chemical traits and for use in poplar breeding programs for cellulosic ethanol production. Integration of NIR analysis in multi-site clonal trials would allow simultaneous multi-trait evaluation and give access to identify potential trade-offs between biomass production and biomass composition, which in turn, supports poplar breeding programs to better monitor multi-trait selection and exploit the large variation present in natural gene pools. As a first check at the genotypic level, we have computed the correlations within each harvest between growth and wood properties and haven't found any adverse correlation within our dataset (Fig. S6). These results are encouraging towards the development of performing clones dedicated to biomass and biofuel production.

Despite its importance, optimal procedures for developing NIR calibrations for rapid prediction of wood composition in multi-site poplar clonal trials are not well established. In the present study, we developed NIR calibration models and successfully applied this indirect method to analyze the sources and extent of variability for wood chemical traits in large-scale clonal trials of *P. nigra*, which is the first work, as far as we know. Finally, future work on development of new calibration models would be useful to further establish the NIR calibration protocols for clonal trials. Some of the important points to consider will be the number of technical replicates for the reference samples to reduce the uncertainties associated with the standard methods and the number of biological replicates per genotype to reach enough accuracy on a clonal basis.

5. Conclusions

From our study of wood chemical traits in clonal trials of European black poplar at two contrasting sites, three important conclusions can

be drawn. (1) We successfully developed global and site specific NIR calibration models for predicting wood chemical traits in natural populations of European black poplar with reasonable accuracy. (2) We demonstrated the high throughput nature of the NIR method, by applying the calibrations to predict the wood chemical composition of the 5799 trees and by the analyses of these NIR predictions to estimate trait variance components and broad-sense heritabilities. (3) We further used the NIR predictions to test and evaluate the extent of $G \times E$ interaction across coppice rotations within a single site as well as across sites.

In this study, the moderate to high heritability estimates and the detection of $G \times E$ interaction suggests that the NIR-based technique can efficiently be used for dissecting the genetic basis of wood chemical properties in a multi-environment large-scale poplar clonal trials and for screening elite individuals from such trials as future parents for interspecific hybridization. Integration of such indirect method in poplar tree breeding programs would allow the exploitation of standing genetic variation in poplars for developing poplar genotypes that combine high biomass yield with superior wood quality for cellulosic ethanol production. Furthermore, the observed moderate to strong genetic control over the NIR-predicted wood chemical traits should pave the way for more detailed dissection of the genetic and molecular basis of the NIR-predicted wood compositional variation through molecular marker analysis of the NIR predictions. In particular, it would be useful to extend such analysis to association mapping aimed at identifying individual loci controlling the predicted phenotypic variation in the studied population of *P. nigra*.

Acknowledgements

The authors gratefully acknowledge the staff of the INRA-GBFOR experimental unit for the establishment and management of the experimental plantation in Orléans, wood sample collections in the two sites, and contribution to circumference measurements; Alasia Franco Vivai staff for the management of the experimental plantation in Savigliano. We would also like to thank Eduardo Cappa, Facundo Muñoz and Leopoldo Sanchez for useful discussion on the genetic analysis of data.

Funding: Establishment and management of the experimental sites until harvests were carried out with financial support from the NOVELTREE project (EU-FP7-211868). All analyses on wood samples were supported by the SYBIOPOP project funded by the French National Research Agency (ANR-13-JSV6-0001). M. N. G. was supported by a PhD grant jointly funded by the SYBIOPOP project (ANR-13-JSV6-0001) and the EFPA division of INRA.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at <http://dx.doi.org/10.1016/j.indcrop.2017.05.013>.

References

- Abramson, M., Shoseyov, O., Shani, Z., 2010. Plant cell wall reconstruction toward improved lignocellulosic production and processability. *Plant Sci.* 178, 61–72. <http://dx.doi.org/10.1016/j.plantsci.2009.11.003>.
- Alves, A., Schwanninger, M., Pereira, H., Rodrigues, J., 2006. Calibration of NIR to assess lignin composition (H/G ratio) in maritime pine wood using analytical pyrolysis as the reference method. *Holzforschung* 60, 29–31. <http://dx.doi.org/10.1515/HF.2006.006>.
- Alves, A., Simões, R., Stackpole, D., Vaillancourt, R., Potts, B., Schwanninger, M., Rodrigues, J., 2011. Determination of the syringyl/guaiacyl (S/G) ratio of *Eucalyptus globulus* Labill. wood lignin by NIR-based PLS-R models using analytical pyrolysis as the reference method. *J. Near Infrared Spectrosc.* 19, 343–348.
- Alves, A.M.M., Simões, R.F.S., Santos, C.A., Potts, B.M., Rodrigues, J., Schwanninger, M., 2012. Determination of *Eucalyptus globulus* wood extractives content by near infrared-based partial least squares regression models: comparison between extraction procedures. *J. Near Infrared Spectrosc.* 20, 275–285. <http://dx.doi.org/10.1255/jnirs.987>.
- Araus, J.L., Cairns, J.E., 2014. Field high-throughput phenotyping: the new crop breeding frontier. *Trends Plant Sci.* 19, 52–61. <http://dx.doi.org/10.1016/j.tplants.2013.09.008>.
- Baillères, H., Davrieux, F., Ham-Pichavant, F., 2002. Near infrared analysis as a tool for rapid screening of some major wood characteristics in a eucalyptus breeding program. *Ann. For. Sci.* 59, 479–490. <http://dx.doi.org/10.1051/forest:2002032>.
- Bradshaw, H.D., Ceulemans, R., Davis, J., Stettler, R., 2000. Emerging model systems in plant biology: poplar (*Populus*) as a model forest tree. *J. Plant Growth Regul.* 19, 306–313. <http://dx.doi.org/10.1007/s003440000030>.
- Cagelli, L., Lefevre, F., 1995. The conservation of *Populus nigra* L. and gene flow with cultivated poplars in Europe. *For. Genet.* 2, 135–144.
- Ceulemans, R., Deraedt, W., 1999. Production physiology and growth potential of poplars under short-rotation forestry culture. *For. Ecol. Manage.* 121, 9–23. [http://dx.doi.org/10.1016/S0378-1127\(98\)00564-7](http://dx.doi.org/10.1016/S0378-1127(98)00564-7).
- Costa e Silva, J., Borralho, N.M.G., Araújo, J.A., Vaillancourt, R.E., Potts, B.M., Silva, J.C.E., Araujo, J.A., 2008. Genetic parameters for growth, wood density and pulp yield in *Eucalyptus globulus*. *Tree Genet. Genomes* 5, 291–305. <http://dx.doi.org/10.1007/s11295-008-0174-9>.
- Da Silva Perez, D., Guillemain, A., Alazard, P., Plomion, C., Rozenberg, P., Carlos Rodrigues, J., Alves, A., Chantre, G., 2007. Improvement of *Pinus pinaster* Ait elite trees selection by combining near infrared spectroscopy and genetic tools. *Holzforstung* 61, 611–622. <http://dx.doi.org/10.1515/HF.2007.118>.
- Dutkowski, G.W., Silva, J.C.E., Gilmour, A.R., Lopez, G.A., 2002. Spatial analysis methods for forest genetic trials. *Can. J. For. Res.* 32, 2201–2214. <http://dx.doi.org/10.1139/x02-111>.
- Frison, E., Lefevre, F., de Vries, S., Turok, J., 1994. *Populus Nigra* Network: Report of the First Meeting 3–5 October 1994 Izmit, Turkey. IPGRI, Rome, Izmit, Turkey.
- Gaspar, M.J., Alves, A., Louzada, J.L., Morais, J., Santos, A., Fernandes, C., Almeida, M.H., Rodrigues, J.C., 2011. Genetic variation of chemical and mechanical traits of maritime pine (*Pinus pinaster* Aiton). Correlations with wood density components. *Ann. For. Sci.* 68, 255–265. <http://dx.doi.org/10.1007/s13595-011-0034-x>.
- Guerra, F.P., Wegrzyn, J.L., Sykes, R., Davis, M.F., Stanton, B.J., Neale, D.B., 2013. Association genetics of chemical wood properties in black poplar (*Populus nigra*). *New Phytol.* 197, 162–176. <http://dx.doi.org/10.1111/nph.12003>.
- Guerra, F.P., Richards, J.H., Fiehn, O., Famula, R., Stanton, B.J., Shuren, R., Sykes, R., Davis, M.F., Neale, D.B., 2016. Analysis of the genetic variation in growth, ecophysiology, and chemical and metabolomic composition of wood of *Populus trichocarpa* provenances. *Tree Genet. Genomes* 12. <http://dx.doi.org/10.1007/s11295-015-0965-8>.
- Guét, J., Fabbri, F., Fichot, R., Sabatti, M., Bastien, C., Brignolas, F., 2015. Genetic variation for leaf morphology, leaf structure and leaf carbon isotope discrimination in European populations of black poplar (*Populus nigra* L.). *Tree Physiol.* 35, 850–863. <http://dx.doi.org/10.1093/treephys/tpv056>.
- Hamilton, M.G., Raymond, C., Harwood, C., Potts, B., 2009. Genetic variation in *Eucalyptus nitens* pulpwood and wood shrinkage traits. *Tree Genet. Genomes* 5, 307–316. <http://dx.doi.org/10.1007/s11295-008-0179-4>.
- Henderson, C.R., 1984. Applications of Linear Models in Animal Breeding Models 384 Univ. Guelph <http://dx.doi.org/10.1002/9780470316856.ch7>.
- Isik, F., Mora, C.R., Schimleck, L.R., 2011. Genetic variation in *Pinus taeda* wood properties predicted using non-destructive techniques. *Ann. For. Sci.* 68, 283–293. <http://dx.doi.org/10.1007/s13595-011-0035-9>.
- Jiang, W., Han, G., Via, B.K., Tu, M., Liu, W., Fasina, O., 2014. Rapid assessment of coniferous biomass lignin-carbohydrates with near-infrared spectroscopy. *Wood Sci. Technol.* 48, 109–122. <http://dx.doi.org/10.1007/s00226-013-0590-3>.
- Kačík, F., Đurkovič, J., Kačíková, D., 2012. Chemical profiles of wood components of poplar clones for their energy utilization. *Energies* 5, 5243–5256. <http://dx.doi.org/10.3390/en5125243>.
- Kennard, R., Stone, L.A., 1969. Computer aided design of experiments. *Technometrics* 11, 137–148.
- Kube, P.D., Raymond, C.A., Banham, P.W., 2001. Genetic parameters for diameter, basic density, cellulose content and fibre properties for *Eucalyptus nitens*. *For. Genet.* 8, 285–294.
- Li, B.B., Morris, J., Martin, E.B., 2002. Model selection for partial least squares regression. *Chemom. Intell. Lab. Syst.* 64, 79–89.
- Li, H., Liang, Y., Xu, Q., Cao, D., 2009. Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration. *Anal. Chim. Acta* 648, 77–84. <http://dx.doi.org/10.1016/j.aca.2009.06.046>.
- Lin, C.-S., Binns, M.R., Lefkovich, L.P., 1986. Stability analysis: where do we stand? *Crop Sci.* 26, 894–900.
- Lupoi, J.S., Singh, S., Simmons, B.A., Henry, R.J., 2014. Assessment of lignocellulosic biomass using analytical spectroscopy: an evolution to high-throughput techniques. *Bioenergy Res.* 7, 1–23. <http://dx.doi.org/10.1007/s12155-013-9352-1>.
- Lynch, M., Walsh, B., 1998. *Genetics and Analysis of Quantitative Traits*. Sinauer Associates Sunderland, MA.
- Mevik, B.H., Wehrens, R., 2007. The pls package: principal component and partial least squares regression in R. *J. Stat. Softw.* 18, 1–23. <http://dx.doi.org/10.1159/000323281>.
- Muñoz, F., Sanchez, L., 2015. *BreedR: Statistical Methods for Forest Genetic Resources Analysts*.
- Muir, W., Nyquist, W.E., Xu, S., 1992. Alternative partitioning of the genotype-by-environment interaction. *Theor. Appl. Genet.* 84, 193–200. <http://dx.doi.org/10.1007/BF00224000>.
- Neale, D.B., Kremer, A., 2011. Forest tree genomics: growing resources and applications. *Nat. Rev. Genet.* 12, 111–122. <http://dx.doi.org/10.1038/nrg2931>.
- Poke, F.S., Potts, B.M., Vaillancourt, R.E., Raymond, C.A., 2006. Genetic parameters for lignin, extractives and decay in *Eucalyptus globulus*. *Ann. For. Sci.* 63, 813–821.


- <http://dx.doi.org/10.1051/forest:2006080>.
- Porth, I., Klápště, J., Skyba, O., Lai, B.S.K., Geraldes, A., Muchero, W., Tuskan, G.A., Douglas, C.J., El-Kassaby, Y.A., Mansfield, S.D., 2013. *Populus trichocarpa* cell wall chemistry and ultrastructure trait variation, genetic control and genetic correlations. *New Phytol.* 197, 777–790. <http://dx.doi.org/10.1111/nph.12014>.
- R Core Team, 2015. R: a Language and Environment for Statistical Computing. R Foundation for Statistical Computing. R Core Team, Vienna, Austria. URL <https://www.R-project.org/>.
- Ragauskas, A.J., Nagy, M., Kim, D.H., Eckert, C.A., Hallett, J.P., Liotta, C.L., 2006. From wood to fuels: integrating biofuels and pulp production. *Ind. Biotechnol.* 2, 55–65. <http://dx.doi.org/10.1089/ind.2006.2.55>.
- Raymond, C.A., Schimleck, L.R., 2002. Development of near infrared reflectance analysis calibrations for estimating genetic parameters for cellulose content in *Eucalyptus globulus*. *Can. J. For. Res. Can. Rech. For.* 32, 170–176. <http://dx.doi.org/10.1139/X01-174>.
- Raymond, C.A., Schimleck, L.R., Muneri, A., Michell, A.J., 2001. Genetic parameters and genotype-by-environment interactions for pulp yield predicted using near infrared reflectance analysis and pulp productivity in *Eucalyptus globulus*. *For. Genet.* 8, 213–224.
- Robinson, A.R., Mansfield, S.D., 2009. Rapid analysis of poplar lignin monomer composition by a streamlined thioacidolysis procedure and near-infrared reflectance-based prediction modeling. *Plant J.* 58, 706–714. <http://dx.doi.org/10.1111/j.1365-3113X.2009.03808.x>.
- Rodrigues, J., Meier, D., Faix, O., Pereira, H., 1999. Determination of tree to tree variation in syringyl/guaiacyl ratio of *Eucalyptus globulus* wood lignin by analytical pyrolysis. *J. Anal. Appl. Pyrolysis* 48, 121–128. [http://dx.doi.org/10.1016/S0165-2370\(98\)00134-X](http://dx.doi.org/10.1016/S0165-2370(98)00134-X).
- Rodrigues, J., Graça, J., Pereira, H., 2001. Influence of tree eccentric growth on syringyl/guaiacyl ratio in *Eucalyptus globulus* wood lignin assessed by analytical pyrolysis. *J. Anal. Appl. Pyrolysis* 58–59, 481–489. [http://dx.doi.org/10.1016/S0165-2370\(00\)00121-2](http://dx.doi.org/10.1016/S0165-2370(00)00121-2).
- Rubin, E.M., Himmel, M.E., Ding, S., Johnson, D.K., Adney, W.S., 2007. Biomass Recalcitrance. *Nat.* 454, 804–807. <http://dx.doi.org/10.1126/science.1137016>.
- Rubin, E.M., 2008. Genomics of cellulosic biofuels. *Nature* 454, 841–845. <http://dx.doi.org/10.1038/nature07190>.
- Sannigrahi, P., Ragauskas, A.J., Tuskan, G.A., 2010. Poplar as a feedstock for biofuels: a review of compositional characteristics. *Biofuels Bioprod. Biorefin.* 4, 209–226. <http://dx.doi.org/10.1002/bbb.206>.
- Schimleck, L.R., Kube, P.D., Raymond, C.A., 2004. Genetic improvement of kraft pulp yield in *Eucalyptus nitens* using cellulose content determined by near infrared spectroscopy. *Can. J. For. Res.* 34, 2362–2370. <http://dx.doi.org/10.1139/X04-119>.
- Schubert, C., 2006. Can biofuels finally take center stage? *Nat. Biotechnol.* 24, 777–784. <http://dx.doi.org/10.1038/nbt0706-777>.
- Schwanninger, M., Rodrigues, J.C., Gierlinger, N., Hinterstoisser, B., 2011a. Determination of lignin content in Norway spruce wood by Fourier transformed near infrared spectroscopy and partial least squares regression. Part 1. Wavenumber-selection and evaluation of the selected range. *J. Near Infrared Spectrosc.* 19, 319–329. <http://dx.doi.org/10.1255/jnirs.945>.
- Schwanninger, M., Rodrigues, J., Hinterstoisser, B., 2011b. Determination of lignin content in Norway spruce wood by Fourier transformed near infrared spectroscopy and partial least squares regression analysis. Part 2: Development and evaluation of the final model. *J. Near Infrared Spectrosc.* 19, 331–341.
- Shelbourne, C.J.A., 1972. Genotype-environment interaction: its study and its implications in forest tree improvement. *Proc. IUFRO Genet.-SABARAO Joint Symp. Tokyo B-1 (I)*, 1–28.
- Stackpole, D.J., Vaillancourt, R.E., Downes, G.M., Harwood, C.E., Potts, B.M., 2010. Genetic control of kraft pulp yield in *Eucalyptus globulus*. *Can. J. For. Res.* 40, 917–927. <http://dx.doi.org/10.1139/X10-035>.
- Stackpole, D.J., Vaillancourt, R.E., Alves, A., Rodrigues, J., Potts, B.M., 2011. Genetic variation in the chemical components of *Eucalyptus globulus* Wood. *G3: Genes Genomes Genet.* G3 (1), 151–159. <http://dx.doi.org/10.1534/g3.111.000372>.
- Stevens, A., Ramirez-Lopez, L., 2013. An introduction to the prospectr package.
- Sticklen, M.B., 2008. Plant genetic engineering for biofuel production: towards affordable cellulosic ethanol. *Nat. Rev. Genet.* 9, 433–443. <http://dx.doi.org/10.1038/nrg2336>.
- Tsuchikawa, S., Kobori, H., 2015. A review of recent application of near infrared spectroscopy to wood science and technology. *J. Wood Sci.* 61, 213–220. <http://dx.doi.org/10.1007/s10086-015-1467-x>.
- Wegrzyn, J.L., Eckert, A.J., Choi, M., Lee, J.M., Stanton, B.J., Sykes, R., Davis, M.F., Tsai, C.J., Neale, D.B., 2010. Association genetics of traits controlling lignin and cellulose biosynthesis in black cottonwood (*Populus trichocarpa*, Salicaceae) secondary xylem. *New Phytol.* 188, 515–532. <http://dx.doi.org/10.1111/j.1469-8137.2010.03415.x>.
- Yang, W., Guo, Z., Huang, C., Duan, L., Chen, G., Jiang, N., Fang, W., Feng, H., Xie, W., Lian, X., Wang, G., Luo, Q., Zhang, Q., Liu, Q., Xiong, L., 2014. Combining high-throughput phenotyping and genome-wide association studies to reveal natural genetic variation in rice. *Nat. Commun.* 5, 5087. <http://dx.doi.org/10.1038/ncomms6087>.
- Yeh, T.F., Yamada, T., Capanema, E., Chang, H.M., Chiang, V., Kadla, J.F., 2005. Rapid screening of wood chemical component variations using transmittance near-infrared spectroscopy. *J. Agric. Food Chem.* 53, 3328–3332. <http://dx.doi.org/10.1021/jf0480647>.
- Zhang, J., Novaes, E., Kirst, M., Peter, G.F., 2014. Comparison of pyrolysis mass spectrometry and near infrared spectroscopy for genetic analysis of lignocellulose chemical composition in *Populus*. *Forests* 5, 466–481. <http://dx.doi.org/10.3390/f5030466>.
- Zhang, P., Wu, F., Kang, X., 2015. Chemical properties of wood are under stronger genetic control than growth traits in *Populus tomentosa* Carr. *Ann. For. Sci.* 72, 89–97. <http://dx.doi.org/10.1007/s13595-014-0401-5>.
- signal developers, 2013. Signal: Signal Processing. URL: <http://r-forge.r-project.org/projects/signal/>.
- Zhou, G., Taylor, G., Polle, A., 2011. FTIR-ATR-based prediction and modelling of lignin and energy contents reveals independent intra-specific variation of these traits in bioenergy poplars. *Plant Methods* 7, 1–10. <http://dx.doi.org/10.1186/1746-4811-7-9>.

RESEARCH ARTICLE

Open Access



Accuracy of RNAseq based SNP discovery and genotyping in *Populus nigra*

Odile Rogier¹, Aurélien Chateigner¹, Souhila Amanzougarene¹, Marie-Claude Lesage-Descauses¹, Sandrine Balzergue^{2,3}, Véronique Brunaud², José Caius², Ludivine Soubigou-Taconnat², Véronique Jorge¹ and Vincent Segura^{1*} 

Abstract

Background: *Populus nigra* is a major tree species of ecological and economic importance for which several initiatives have been set up to create genomic resources. In order to access the large number of Single Nucleotide Polymorphisms (SNPs) typically needed to carry out a genome scan, the present study aimed at evaluating RNA sequencing as a tool to discover and type SNPs in genes within natural populations of *P. nigra*.

Results: We have devised a bioinformatics pipeline to call and type SNPs from RNAseq reads and applied it to *P. nigra* transcriptomic data. The accuracy of the resulting RNAseq-based SNP calling and typing has been evaluated by (i) comparing their position and alleles to those previously reported in candidate genes, (ii) assessing their genotyping accuracy with respect to a previously available SNP chip and (iii) evaluating their inter-annual repeatability. We found that a combination of several callers yields a good compromise between the number of variants type and the accuracy of genotyping. We further used the resulting genotypic data to carry out basic genetic analyses whose results confirm the quality of the RNAseq-based SNP dataset.

Conclusions: We demonstrated the potential and accuracy of RNAseq as an efficient way to genotype SNPs in *P. nigra*. These results open prospects towards the use of this technology for quantitative and population genomics studies.

Keywords: DNA polymorphisms, Bioinformatics pipeline, Black poplar, Transcriptomics

Background

Populus nigra is a major tree species from Eurasian riparian ecosystems and one of the 3 main parental species used in poplar breeding programs to develop highly productive interspecific cultivated hybrids. For these reasons, several initiatives have recently been set up to create genomic resources within this species as tools to improve conservation and breeding strategies [1, 2]. The main objective of such initiatives is to discover and type genomic variants like Single Nucleotide Polymorphisms (SNPs) for various applications, including the identification and quantification of introgressions from the cultivated compartment, the study of population structure and the identification of variants associated with economically or ecologically relevant phenotypes through association genetics.

Early studies in *P. nigra* have focused on re-sequencing specific candidate genes from the lignin pathway [3–5], but more recent work has broadened the scope of analyses through the development of a genotyping chip from SNPs detected by whole-genome sequencing [1, 2]. This genotyping tool was successfully used to study the structure of the genetic diversity of the species [1] and to identify some genomic regions associated with economically important traits [6]. However, the genotyping was limited to 7903 SNPs preferentially located within particular candidate regions underlying some Quantitative Trait Loci (QTLs) previously reported in biparental crosses. Moreover, the frequency of the SNPs within *P. nigra* populations appeared to be upwardly biased, limiting the analyses to common variants [1]. Consequently, the application of this chip especially in association genetics could be limited as underlined by the low number of significant associations reported [6]. Indeed, given the rapid Link-

*Correspondence: vincent.segura@inra.fr

¹BioForA, INRA, ONF, 45075 Orléans, France

Full list of author information is available at the end of the article



age Disequilibrium (LD) decay within this species and its genome size, an exhaustive genome-wide association study (GWAS) would require between 67,000 and 134,000 evenly spaced SNPs which is between 8 and 16 times more than the number of SNPs available from the chip cited above [7, 8].

In order to access a large number of SNPs, as typically needed for an exhaustive GWAS in *P. nigra*, several options relying on next-generation sequencing would be available. If whole genome sequencing appears to be still too expensive for a fairly large number of genotypes, reducing the complexity of the genome prior to sequencing for instance with restriction enzymes (GBS [9]; RADseq [10]), or sequence capture (exome sequencing, [11]) seems to be a promising way forward for reaching the objectives. Indeed, sequence capture has recently successfully been used to genotype around 350,000 SNPs in *P. deltoides* and identify putative regulators of bioenergy traits [12]. RNA sequencing (RNAseq) represents also a cost-effective way to reduce complexity while focusing on the expressed fraction of the genome [13]. However, to date, RNAseq has more often been used for SNP discovery than for direct genotyping of large populations. For instance, Geraldès et al. [14] found around 500,000 SNPs through RNAseq of developing secondary xylem in *P. trichocarpa*, and later on, a SNP chip was developed partly from the previously discovered RNAseq SNPs [8] in order to further carry out association scans [15, 16]. Nevertheless, recent studies have been using RNAseq as a tool for both discovering and genotyping a large number of SNPs in populations [17–21], underlining the interest of this approach for population and quantitative genomics studies. However, to our knowledge, no study so far has evaluated the accuracy of SNP genotyping from RNAseq data.

The present study aims at evaluating RNAseq as a tool to type a sufficiently large amount of SNPs within natural populations of *P. nigra* to carry out a GWAS. For that purpose, we performed RNAseq on pools of young differentiated xylem and cambium collected on 2 biological replicates of 12 genotypes originated from 6 natural populations. We have further developed a dedicated bioinformatic pipeline to discover and type SNPs within the sequences. The accuracy of the resulting RNAseq-based SNPs has also been evaluated by (i) comparing their position and alleles to those previously reported in candidate genes [3, 4], (ii) assessing their genotyping accuracy with respect to a SNP chip [1], (iii) evaluating their interannual repeatability. Finally, the resulting validated SNPs have been used to perform basic genetic analyses to illustrate the usefulness of the released SNP dataset.

Methods

Plant material, experimental design and tissue sampling

Trees were sampled in an experimental site established in a common garden in 2008 in Central France (Orléans, Loire Valley, 47°50'N 01°54'E, 108 m above sea level) at INRA. The experimental site is described in Guet et al. [22]. Briefly, a *P. nigra* collection composed of 1098 cloned genotypes sampled in natural populations present in 11 river catchments in four European countries was planted according to a randomized complete block design with a single tree per block and six replicates per genotype. The trees have been growing through three short rotations since the planting, they were cut back in March 2010 and in February 2012. The experiment was carried out in accordance with local legislation.

Twelve genotypes belonging to 6 river catchments (as defined by Guet et al. [22]: Adour, Dranse, Loire, Ramières, Rhin, Ticino) were selected for the present study to represent the range of available geographical origin in France and Northern Italy. The genotypes from the French populations (Adour: BDX-003, AST-005; Dranse: DRA-045, DRA-038; Loire: VDL-018, 92510-1; Ramières: 1-J31, 1-A26; and Rhin: STR-010, RHN-028) were collected and are owned by INRA (UMR0588-BioForA), while those from Italy (Ticino: SN-2, SN-7) are owned and were kindly provided by the University of Tuscia. Two trees per genotype were sampled in June 2014 (in blocks 2 and 4). The most vigorous stem of each tree was cut back and the bark was detached from the trunk in order to scratch young differentiating xylem and cambium tissues using a scalpel. The tissues were immediately immersed in liquid nitrogen and crudely ground before storage at -80°C pending the RNA extraction.

RNA extraction, library preparation and sequencing

For each biological repetition and each tissue, samples of young differentiating xylem and cambium were ground with a swing mill (Retsch, Germany) and tungsten beads under cryogenic conditions with liquid nitrogen during 25 s (frequency 25 cps/s). Powders were stored at -80°C until RNA extraction. About 100 mg of ground tissue was used to isolate separately total RNA from xylem and cambium of each plant with RNeasy Plant kit (Qiagen, France) according to manufacturer's recommendations. Treatment with DNase I (Qiagen, France) to ensure elimination of genomic DNA was made during this purification step. RNA was eluted in RNase-DNase free water and quantified with a Nanodrop spectrophotometer. RNA from xylem and cambium of the same plant were pooled in an equimolar extract (250 ng/ μ L) before being sent to the sequencing platform.

RNAseq experiment was carried out at the platform POPS (transcriptOmic Platform of Institute of Plant Sciences - Paris-Saclay) thanks to IG-CNS Illumina HiSeq2000. RNAseq libraries were prepared from polyA RNA selection using TruSeq Stranded mRNA SamplePrep_Guide_15031047_D protocol (Illumina®, California, U.S.A.). Eight libraries were multiplexed per lane and paired-end (PE) sequenced on an Illumina HiSeq2000. Thus, over 22 million of 100 base pairs (bp) PE-reads were generated per sample.

Sequencing data processing and variant calling pipeline

We have devised a bioinformatic pipeline for processing the reads, mapping them to a reference genome and calling the SNPs using several callers (Fig. 1). Each step of this pipeline is described hereafter.

Read quality control was assessed using FastQC (v0.11.4; [23]). Cutadapt 1.10 [24] and the FASTX-toolkit 0.0.13 [25] were used to remove adaptor sequences and low-quality bases. The 13 first 5' bases were removed, as well as bases with PHRED score below 20 from the 3' end of the read. Only reads longer than 35 nucleotides were kept.

Reads were aligned to the *Populus trichocarpa* reference genome version 3.0, retrieved from the JGI Comparative Plant Genomics Portal [26, 27]. Alignment was performed using the short read aligner BWA-MEM 0.7.12 [28] with default parameters using the paired-ends information to produce per-tree SAM files that were converted to BAM files and sorted by aligned position on the reference with SAMtools 1.3 [29]. As an alternative we also tested TopHat [30], but BWA-MEM with default settings yielded the highest percentage of mapping and was thus selected.

The data pre-processing steps recommended in the GATK best practices workflow [31, 32] were performed before variant identification. PCR duplicates were marked with the MarkDuplicates from Picard tools 2.0.1 utility [33] to mitigate biases introduced by data generation steps such as PCR amplification or minimize gene expression variations. We also performed local realignment around indels, checked intron-exon junctions and recalibrated the base quality scores with GATK 3.1 [34].

Four different variant callers were used to perform SNP and indel discovery and genotyping across all 24 samples simultaneously (Table 1): (i) GATK 3.1 [31, 32, 34] using the HaplotypeCaller tool in multi-sample calling mode (modality “GATK”); (ii) GATK 3.1 using the HaplotypeCaller tool in single-sample calling mode followed by joint genotyping of the samples with the GenotypeGVCFs tool (modality “gVCF_GATK”); (iii) FreeBayes 0.9.20 [35] in multi-sample calling mode (modality “FreeBayes”); and (iv) the mpileup command from SAMtools 1.3 [29] in multi-sample calling mode followed by bcftools 1.3.1 [36]

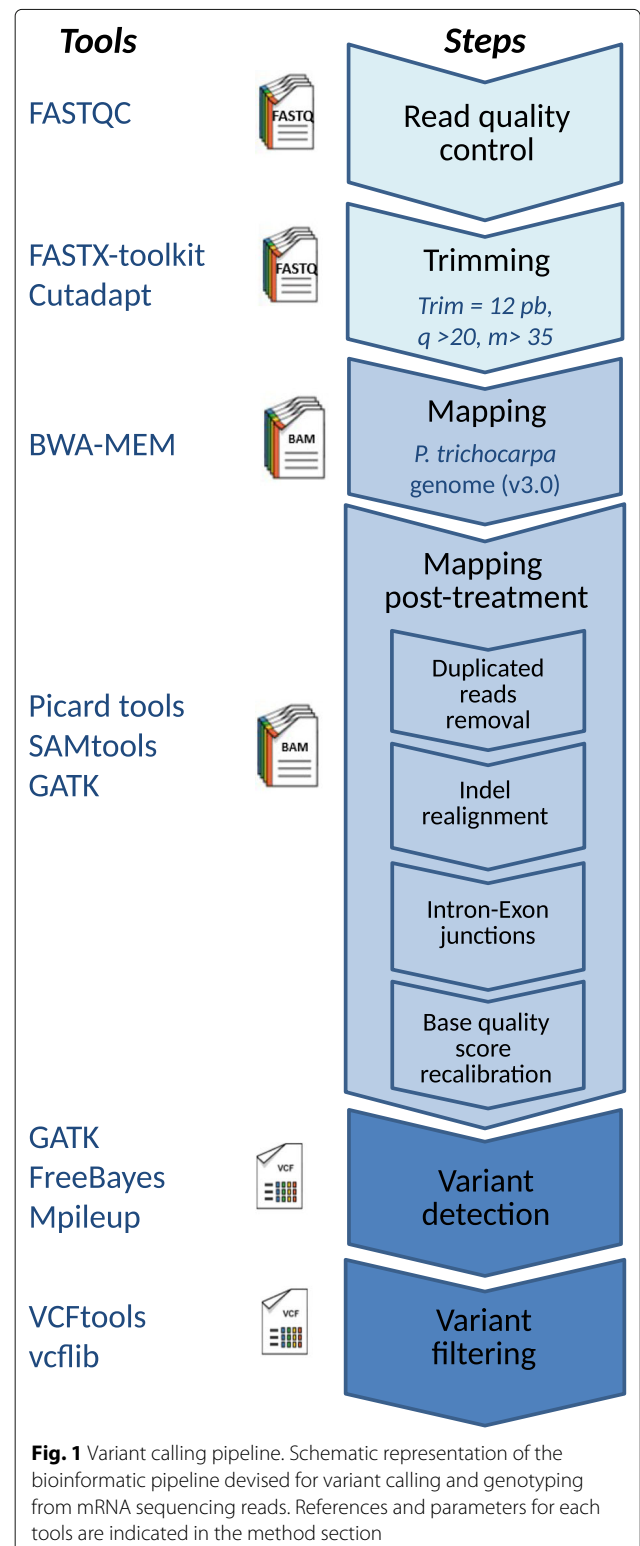


Fig. 1 Variant calling pipeline. Schematic representation of the bioinformatic pipeline devised for variant calling and genotyping from mRNA sequencing reads. References and parameters for each tools are indicated in the method section

with the multiallelic calling model (modality “Mpileup”). Default parameters were used.

We obtained 4 files in a raw Variant Call Format (VCF) with no filter. The functions vcfallelicprimitives

Table 1 Number of variants detected for each of the 7 calling modalities tested in the present study

Variant calling modality	Missing value in genotype calls		
	noNA	2NA	anyNA
GATK	464,829	555,828	902,256
gVCF_GATK	407,037	497,314	927,522
FreeBayes	492,073	640,445	795,459
Mpileup	496,688	594,932	949,411
3Callers	341,584	366,123	400,392
4Callers	252,887	262,447	271,399
3CallersConsensus	356,275	442,931	785,377

noNA: no missing value; *2NA*: lower or equal to 2 missing values; *anyNA*: no filter on missing values. *3Callers*: SNPs detected with at least 3 callers; *4Callers*: SNPs detected with 4 callers; *3CallersConsensus*: SNPs detected with at least 3 callers with correction of the genotype calling when discrepancies existed between callers (see details in Methods)

and `vcfcreatemulti` from `vcflib` [37] were used to decompose the complex variants generated by FreeBayes into a canonical SNP and indel representation. For each caller individually, several filtering parameters were applied with `VCftools` 0.1.15 [38]: selection of biallelic SNPs (indels and not-biallelic SNPs were removed); SNP quality threshold ≥ 30 ; intra-specific polymorphisms (*P. nigra*).

In order to generate a high-confidence SNP set, the SNPs identified by 3 or 4 callers were selected using the `vcf-isec` tool from `VCftools` 0.1.15. We first considered positions with the same genotype across all individuals with 3 or 4 callers (modalities “3Callers” and “4Callers”, Table 1). Because some SNPs could display a difference between callers for only a limited number of individuals, we further considered a consensus set between 3 callers (modality “3CallersConsensus”, Table 1). In this case, for a given individual, when at least 3 callers agreed, the resulting genotype call was set as the consensus between them, otherwise the genotype call was set as a missing value for this particular individual. This part was done using home-made scripts.

In the end, we considered 7 modalities for which we tested 3 different filters for missing values in the genotype calls (Table 1): no missing value allowed (“noNA”), up to 2 missing values allowed (“2NA”) and any missing value allowed (“anyNA”).

Validation of the SNPs detected and genotype calls

A first validation of the SNPs detected and genotyped by each or by combinations of the callers has been done through a comparison of the genotype calls with those previously obtained with a 12k Illumina Infinium Bead-Chip array [1]. Full details of SNP discovery and selection, array development and data filtering criteria are given in Faivre-Rampant et al. [1]. In brief, 852 unrelated *P. nigra* accessions (including our 12 genotypes) were successfully

genotyped with this genotyping array, yielding 7903 SNPs for the validation of genotype calls. For each of the 12 genotypes, genotyping accuracy was calculated as the percentage of similarity between chip genotype and RNAseq genotype at the common positions.

A second validation consisted in comparing the SNPs detected with those previously identified within 5 candidate genes through Sanger and Next-Generation sequencing (CAD4, HCT1, C3H3, CCR7, and 4CL3; [3, 4]). The originally reported SNP were repositioned by aligning reference sequences with the latest *P. trichocarpa* reference genome assembly (v3.0; [26, 27]; Additional file 1).

A third validation consisted in evaluating an inter-annual repeatability of the RNAseq genotype calls by conducting the same experiment on other ramets of the same 12 genotypes one year later. Two other ramets of each genotype were sampled in June 2015 (in blocks 1 and 3 of the same experimental design). The RNA extraction and library preparation were the same as described above. The RNAseq samples have been sequenced in Single-Read (SR) in this second experiment, multiplexing ten samples per lane. This setup yielded approximately 20 millions of SRs per sample. The same bioinformatic pipeline was used on this data, except for the mapping step where we accounted for the single nature of the reads.

The usefulness and relevance of the resulting SNPs for basic genetic studies were further assessed as another form of validation. Minor Allele Frequency (MAF) was calculated with `VCftools` 0.1.15 [38]. Genome-wide distribution of SNPs was calculated based on a 100-kb window with custom R scripts [39]. SNP density within a 100-kb window was further correlated with the sum of the expression of the genes located in the same window. The SNPs have also been annotated using Annovar (version 2017Jul16) [40]. We further tested whether the gene models (with at least 5 SNPs) displayed any enrichment in Gene Ontology (GO) terms using *Arabidopsis thaliana* annotation with the R package `topReviGO` [41]. Finally, population structure was described using a hierarchical ascendant clustering on a distance matrix estimated as $d = 1 - IBS$, where IBS is the identity by state matrix between genotypes computed with PLINK 1.07 [42].

Results

Quality control, mapping and post-treatment

The trimming process removed 0.3% of reads and only 7% of duplicated reads were rejected. At the mapping step, around 99.7% of the reads were mapped against the reference genome (*P. trichocarpa*) and 93.3% were mapped without ambiguous position, even with RNA extracts from a different species (*P. nigra*). A first crude SNP detection and calling on each of the 24 samples with a single caller (“FreeBayes”) enabled the identification of between 772,043 and 1,156,297 SNPs depending on the

sample, of which some were previously genotyped on the same individuals with a SNP array [1]. These common SNPs were used to compare the genotype calls and further identified that 3 of the 24 samples did not match perfectly the original genotypes (genotyping accuracy less than 90%, Additional file 3: Figure S1). These three samples corresponded to one repetition of the 3 genotypes “1-A26”, “RHN-28”, and “STR-10”. They were removed from further analyses. The remaining 21 mapping BAM files were used for SNP detection and genotyping at the genotype level (using genotype as a read group). In other terms, for 9 out of 12 genotypes we used reads from two samples, increasing the sequencing depth available. Validation of genotype identity has been made afterwards using the same SNP array as previously (Additional file 3: Figure S2).

SNP detection and genotyping in 12 genotypes

Between 2,658,024 variants (included intra- and interspecific SNPs and Indels) for “gVCF_GATK” and 3,500,381 variants for “Mpileup” were detected depending on the caller used (Fig. 2; Additional file 2: Table S1). Among filters applied, the selection of *P. nigra* intra-specific SNPs was the criteria that reduced most drastically the numbers of detected SNPs (from 795,459 SNPs for “FreeBayes” to 949,411 SNPs for “Mpileup”). The final number of SNPs detected without missing genotype was fairly similar for all callers, ranging between 407,037 SNPs for “gVCF_GATK” to 496,688 for “Mpileup”.

We further compared the *P. nigra* SNP positions and genotype calls between each callers as well as combinations of at least two callers (Table 1). As expected the number of SNPs detected was lower when considering combinations of callers rather than single ones. Indeed, the “core” SNP set detected by all callers contained 252,887 SNPs with no missing genotype calls (“4Callers-noNA”). This number increased to 341,584 SNPs with no missing genotypes when considering at least 3 callers (“3Callers-noNA”). A further increase could be obtained when computing a consensus genotypes between the callers (“3CallersConsensus-noNA”, 356,275 SNPs) but this gain was much more pronounced when allowing missing genotype calls (“3CallersConsensus-2NA”, 442,931 SNPs; “3CallersConsensus-anyNA”, 785,377 SNPs), underlining the interest of computing a consensus genotyping when combining multiple callers.

SNP validation

A total of 7903 SNPs previously genotyped with a SNP array [1] were compared with the list of SNPs detected with each caller, combination of 3 or 4 callers (Fig. 3; Additional file 2: Table S2; Additional file 3: Figure S3). Genotyping accuracy, evaluated as the percentage of similarity over all common positions, varied from 90 to 99%

and was negatively correlated with the total number of SNP detected and consequently the number of positions available for the comparison. Thus, there is a trade-off between the number of SNPs we are willing to obtain and the quality of the genotyping information. The negative relationship between the number of SNPs detected and the genotyping accuracy appeared to be linear ($R^2 = 0.97$).

The position of the calling methods with respect to the regression line provides information on their performance for variant detection and genotyping. “FreeBayes” and “gVCF_GATK” were always below the line and thus appeared to be the less accurate with respect to the number of variants detected. “GATK” and the combination of 4 callers were always very close to the line and thus could be seen as intermediary performing calling methods. Finally, “Mpileup” and the combination of 3 callers were always above the line, suggesting that they performed best. Of note “3CallersConsensus-anyNA” was the most distant modality above the line, underlining the strength of this approach for detecting and genotyping variants in our dataset.

For further analyses and validations, we decided to focus on the set of SNPs that gave the highest number of SNPs with at least 98% of accuracy, *i.e.* the consensus from 3 callers with no missing data (modality “3CallersConsensus-noNA”). The resulting 356,275 SNP positions were further compared to previously reported *P. nigra* SNPs obtained by Sanger or NGS sequencing of five candidate genes fragments, which were also used to compare detected SNP positions [3, 4] (Fig. 4; Additional file 3: Figure S4). Because these candidate genes were expressed within our samples, a fairly large amount of previously identified SNPs were also detected in our study even within introns. Indeed, the number of positions also detected with RNAseq varied between 30 to 61%. It is worth noting that a fairly large number of SNPs were detected in introns in these candidate genes (“HCT” : 53/70; “4CL3”: 15/36; “C3H3”: 24/45; “CAD4”: 10/19; “CCR7”: 20/39). Another gene was analyzed in detail because it included a large number of SNPs that have been genotyped with the SNP array (Potri.017G084100): 9 from the 19 SNPs used in the array were detected with RNAseq (Additional file 3: Figure S4).

Finally, we carried out an inter-annual repeatability analysis of the genotyping by RNA sequencing approach. Because the second sequencing experiment was done in a single read setting, we detected around twice less positions than in the first experiment (157,569 *vs.* 356,275 with the “3CallersConsensus-noNA” modality). Of note, 88% of the SNPs detected in the second experiment were also found at the same position with the same genotype calls in the first experiment.

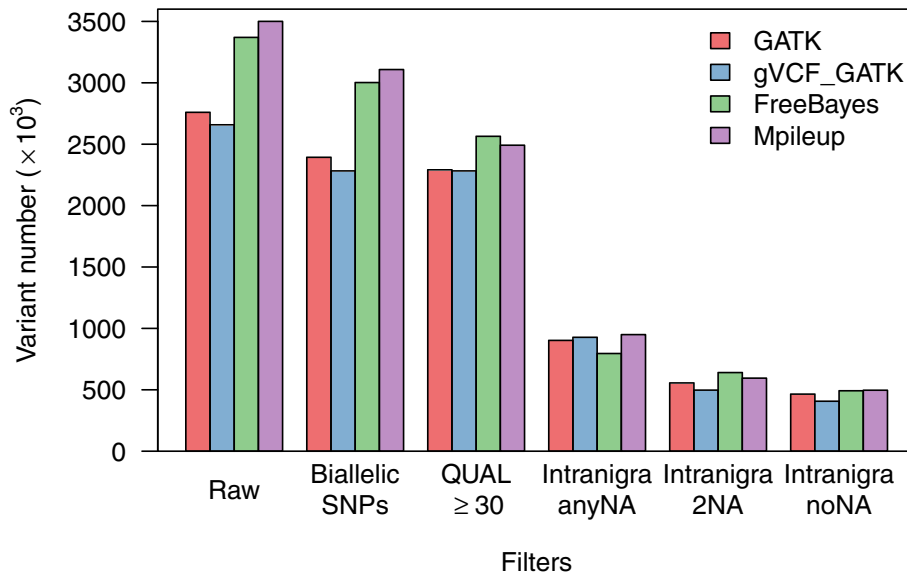


Fig. 2 Variant discovery in 12 *Populus nigra* genotypes. Number of variants discovered with each of the four callers (“GATK” in red; “gVCF_GATK” in blue; “FreeBayes” in green and “Mpileup” in purple) after applying different filters (“Raw”: no filters; “Biallelic SNPs”: indel removed; only biallelic SNPs retained; “QUAL ≥ 30”: SNP quality greater than 30 retained; “Intranigra/anyNA”: SNP polymorphic in *P. nigra* retained; “Intranigra/2NA”: SNP polymorphic in *P. nigra* with at most 2 missing genotype values retained; “noNA”: SNP polymorphic in *P. nigra* without missing genotype value retained)

SNP characterization and usefulness

We estimated the minor allele frequency for each of the 356,275 SNPs from the modality “3CallersConsensus-noNA”. The distribution had an L-shape with an excess of rare alleles as expected under population genetics models (Fig. 5a).

To evaluate the genomic distribution of our SNPs, we computed the density within 100-kb windows of the 351,157 SNPs located on the 19 chromosomes of *P. trichocarpa* v3.0. The number of SNPs within 100-kb windows ranged between 0 and 482 with an average of 89 and a median of 83. Moreover, 92% of the 100-kb

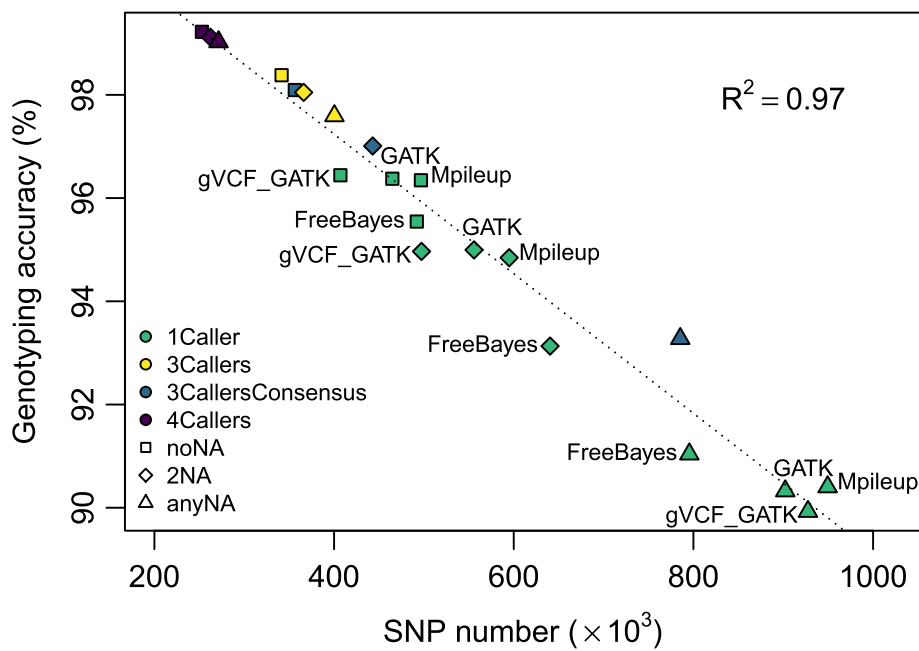


Fig. 3 Relationship between genotyping accuracy and the number of SNPs detected. Number of SNPs detected and genotyping accuracy for 7 calling modalities times 3 options for missing values. See Table 1 for the corresponding denominations

windows harbored at least 1 SNP, underlining an overall good coverage of the genome (Fig. 5b). To further explain the observed variations in SNP density within 100-kb windows, we compared this numbers to the sum of gene expressions within the same windows, estimated as the log₂ of read counts per million. We found a highly significant positive relationship between gene expression and the number of SNPs detected ($R^2 = 0.75$, Fig. 5c).

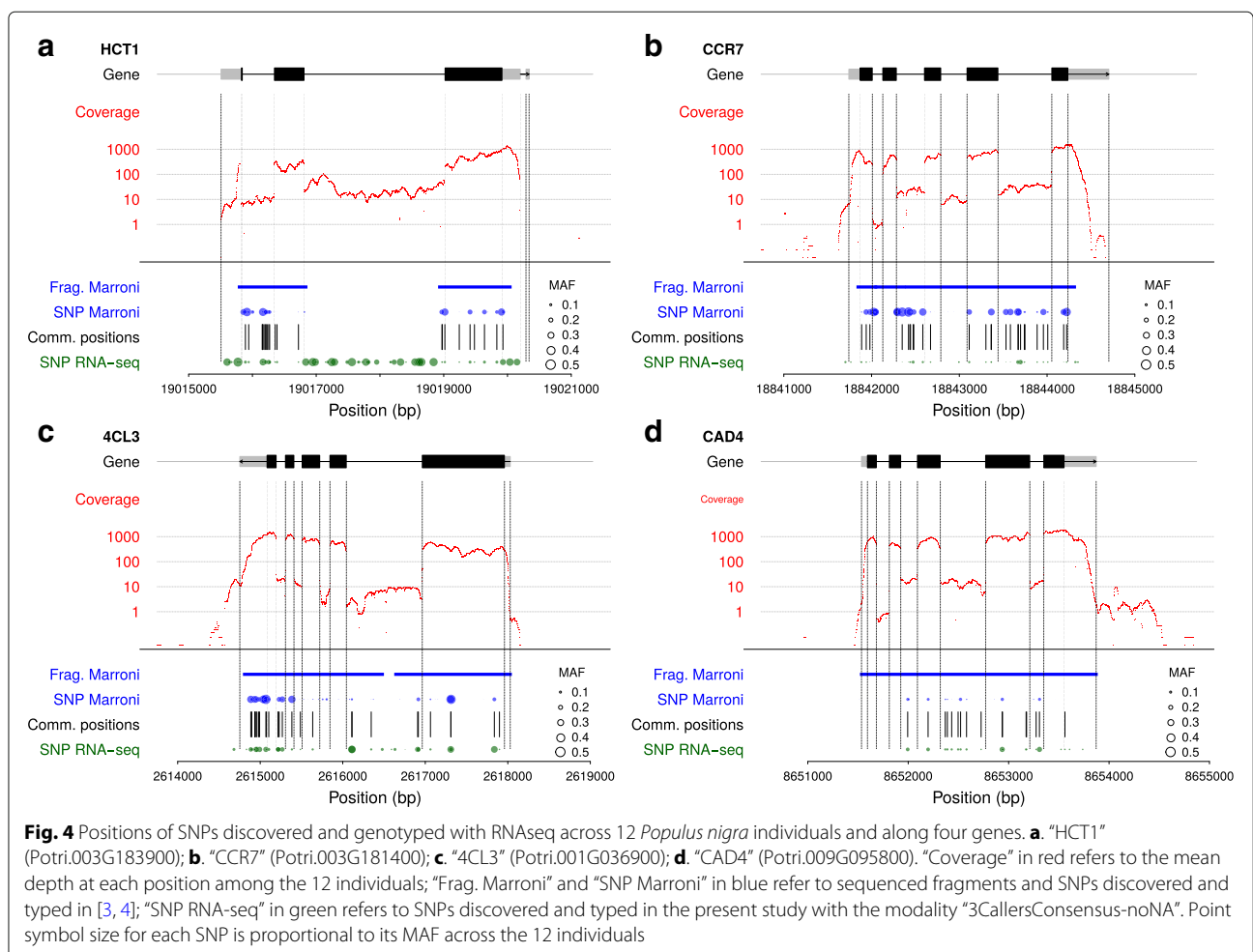
The automatic annotation of our SNPs highlighted as expected that the vast majority of them (80%) were located within exons or 3' and 5' UTRs (Fig. 5d). Nevertheless, as already observed when comparing with SNPs previously reported within candidate genes, a fairly large amount of SNPs were located within introns (15%, Fig. 5d). These intronic SNPs are likely to come from pre-mRNA [43]. Considering exonic SNPs, their annotation highlighted a very low number of mutations affecting the stop codon (1%). The remaining exonic SNPs were almost equally split between synonymous and nonsynonymous sites (Fig. 5d).

In the end, we found that 19,249 genes were covered by at least 5 SNPs which corresponds to 47% of gene models in *P. trichocarpa* genome annotation. We further tested whether these 19,249 gene models displayed any enrichment in GO terms using *Arabidopsis thaliana* annotation (18,384 orthologs). We found that few GO terms were enriched within our set, but they corresponded to biological processes that seem to be quite generic rather than specific to the tissues sampled (Additional file 3: Figure S5).

Finally, we used the 250,784 SNPs with a MAF higher than 5% to evaluate the genetic structure of our 12 genotypes (Fig. 6). A hierarchical ascendant clustering of the genotypes clearly highlighted 6 groups corresponding to the populations to which the genotypes belong to. It is worth mentioning that the population clustering matched their geographic origins.

Discussion

We have successfully built a pipeline from multiple bioinformatics tools for detecting and typing several



hundred thousand SNPs from RNAseq data. Genotyping accuracy of the resulting SNPs has been evaluated by (i) a comparison with genotyping data previously obtained with a SNP array [1] and (ii) an interannual validation. The high accuracy (around 95%) underlined the quality of the genotypic dataset obtained with our pipeline. Additionally, when looking at candidate genes for wood properties (lignin pathway), many SNPs previously reported by DNA sequencing could be recovered within our RNAseq data even if our study focused only on 12 genotypes. This could be expected because 3 and 7 of our 12 genotypes were also included in previous sequencing studies [3, 4]. The resulting variants frequency spectrum followed the expectations from population genetics models and they were spread across most of the genome. The very few genomic regions that appeared to be uncovered correspond to predicted centromeric regions [2] which do not carry many gene models and thus cannot be tagged in an RNAseq experiment.

If the vast majority of SNPs were as expected located within exonic or UTR regions, it is interesting to note that a fairly large number of SNPs appeared to locate within introns. Several hypotheses could explain this result. First, we have used as a reference the genomic annotation from a different species within the same genus: *P. trichocarpa*.

If most of our reads mapped to this reference genome, interspecific variability is likely to have affected the quality of the annotation of our SNPs. In addition, alternative splicing has been shown to be frequent in developing xylem of *P. trichocarpa* [44]. This phenomenon is likely to be more frequent at the interspecific level and may thus have contributed here to the intronic SNPs detected. Second, if in RNAseq most of the reads come from mature mRNA, it has been shown that pre-RNA could as well be sequenced which would yield reads outside of the exonic and UTR regions [43]. Actually, the read coverage was not null in the introns of our candidate genes, providing sufficient information for SNP detection and typing. Thus, it was not surprising to have intronic SNPs within RNAseq data, especially for highly expressed genes as expected here for candidate genes from the lignin pathways, since we sampled our RNA from young differentiating xylem and cambium. Indeed, we have also found a highly significant positive correlation between gene expression level and SNP density, but this observation did not necessarily pop up when we observed the SNPs detected, their frequency and the read coverage on the candidate genes from the lignin pathway. As a matter of fact, the frequency of SNPs did not seem to vary a lot between highly covered exonic regions and weakly covered intronic

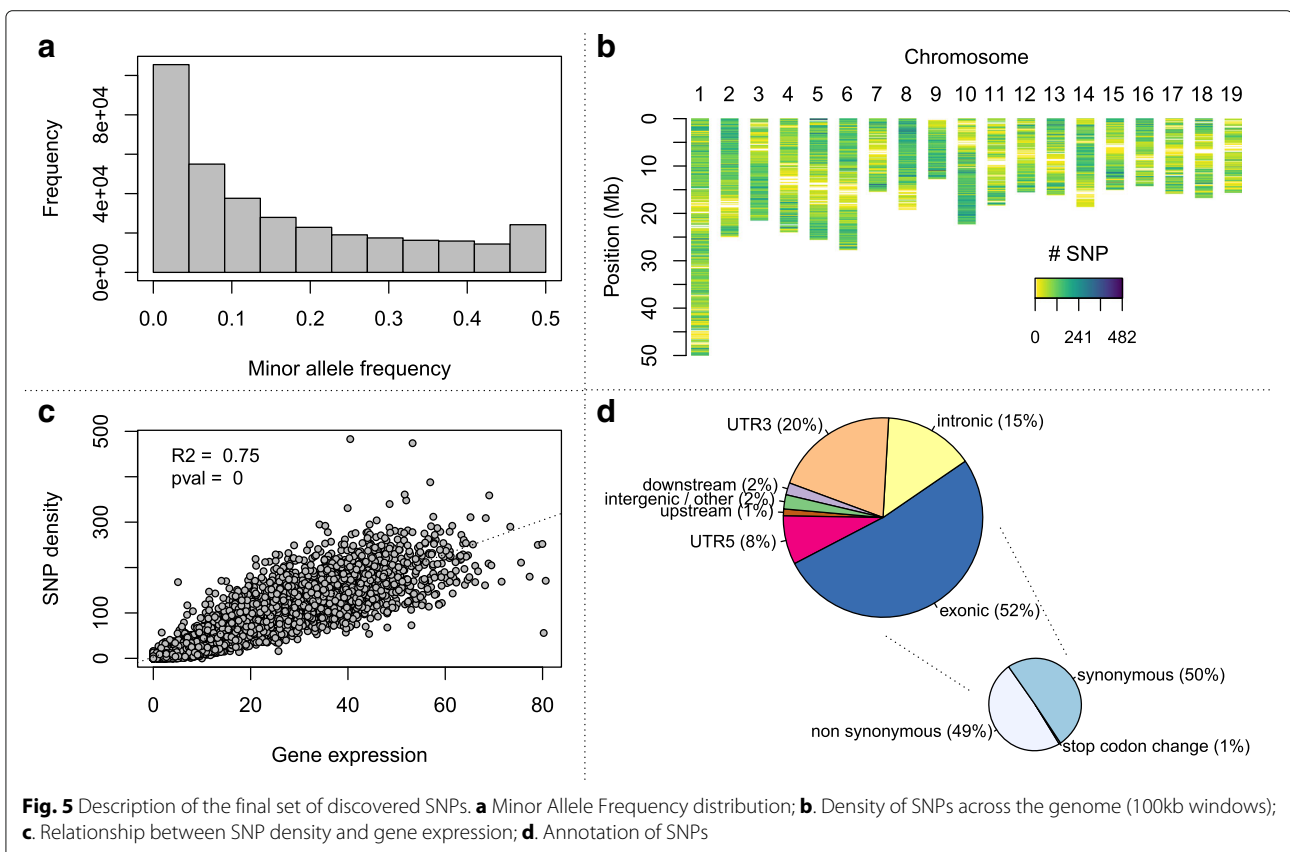


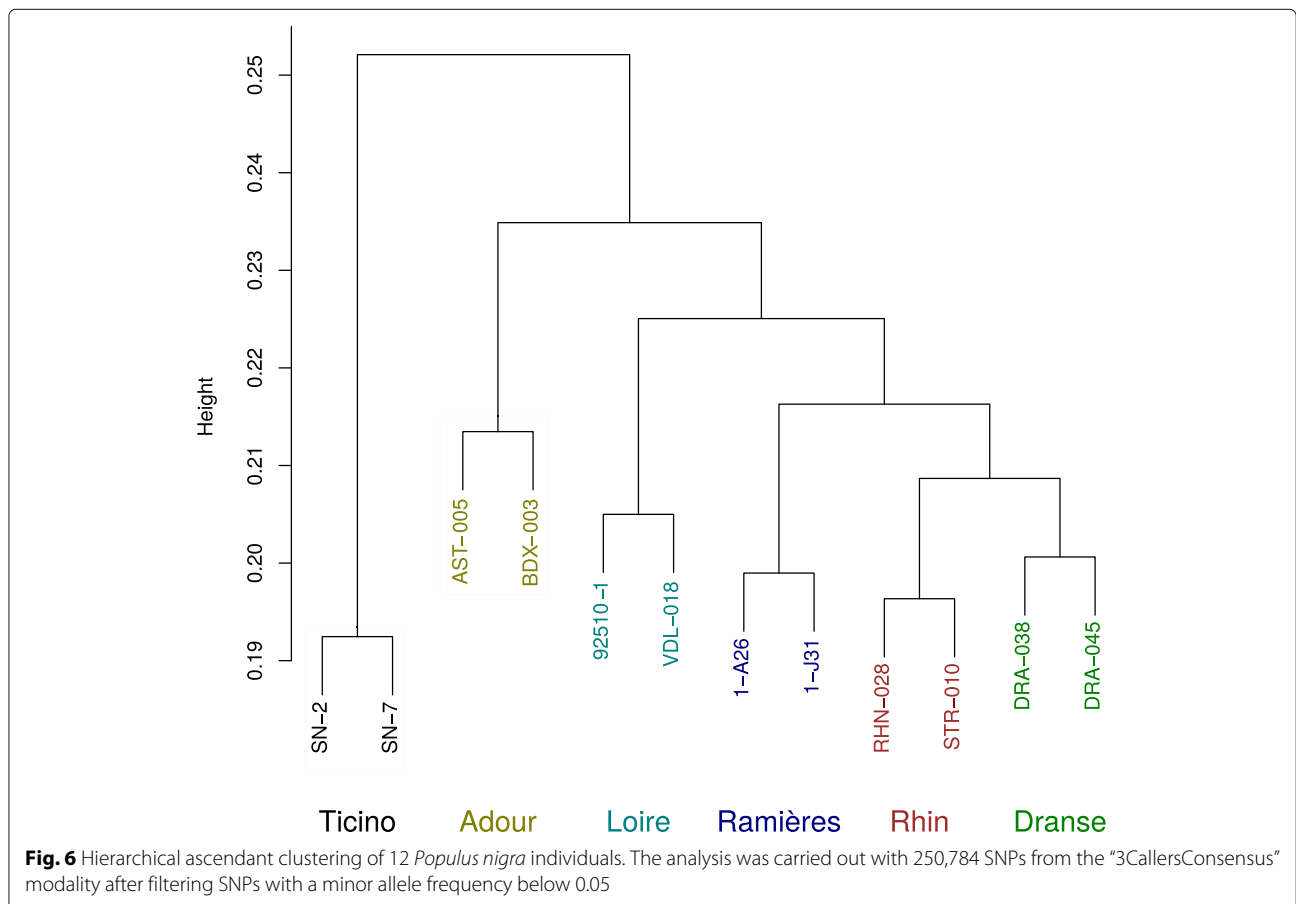
Fig. 5 Description of the final set of discovered SNPs. **a** Minor Allele Frequency distribution; **b**. Density of SNPs across the genome (100kb windows); **c**. Relationship between SNP density and gene expression; **d**. Annotation of SNPs

regions. For “HCT1” (Potri.003G183900), the frequency of SNPs even seems to be higher in introns than exons. This is consistent with the negative Tajima’s D previously obtained on the set of candidate genes from the lignin pathway in *P. nigra* [3, 4] as well as for genes associated with a lower lignin content in *P. deltooides* [12]. In addition, Marroni et al. [3] also reported for “HCT1” a non-synonymous to synonymous nucleotide diversity ratio of 0.03 suggesting that this gene is under purifying selection which may explain the pattern observed here. Consequently, the frequency spectrum of SNPs from RNAseq reads is likely to be complex as both affected by gene expression levels as well as evolutionary factors. Care must thus be taken when using these data for population genomic analyses and especially for detecting signatures of selection.

We used several variant callers as well as their combination. Given the observed and expected trade-off between genotyping accuracy and the number of SNPs detected, we found that this strategy was efficient since better performances could be reached through the combination of multiple callers rather than using a single one, except for Mpileup. The gain was mainly due to the

production of consensus genotype calls from the different callers, especially when they did not all agree. This opens the choice between various options along the accuracy amount trade-off, which could further be picked depending on the objectives of downstream analyses. For instance, if one wishes to obtain the largest number of SNPs within its dataset for carrying out a GWAS, it may be a good idea to use the combination of 3 callers with missing data allowed and then to impute the missing data with a dedicated tool which make use of linkage disequilibrium between neighbouring SNPs for the imputation [45]. If one wants to only use 1 caller, we recommend the use of “Mpileup”, as it is the only one that produced a data set at the equilibrium between quantity and quality of SNPs. Finally, if one wants to have the best quality of SNPs at the price of a lower number, we recommend the use of the 4 callers data intersection, without missing data.

In the present work, we have focused on biallelic SNPs because they constitute the most abundant polymorphism in the genome. However, the callers used have also detected numerous indels or triallelic SNPs which could prove useful for various analyses and thus would deserve



further work. Also, because we used a different species as a reference for mapping our reads and annotating our SNPs, a large amount of the SNPs detected by each of the callers displayed interspecific variation as underlined by the steep decrease in the number of variants when considering intra-nigra polymorphism only. These polymorphisms could also be valuable for species determination and for studying interspecific hybridization [46].

We sampled our RNA from two tissues, young differentiating xylem and cambium, because our research focuses on wood production. Combining information from two tissues has likely increased the number of genes covered by reads and consequently by SNPs compared to what would have been obtained when considering a single tissue. Using this strategy, we could obtain a genotyping dataset with almost half of the gene models of *P. trichocarpa* covered by at least 5 SNPs. Moreover, the GO enrichment analysis suggested that sampling did not introduce a strong bias into the representativeness of functional categories that were effectively captured by the RNAseq experiment. One strategy to increase the genomic coverage could be to combine RNA from multiple tissues but this would have a cost in term of sequencing. More generally, because many factors affect gene expression such as the developmental stage or the tissue considered, further works are required to assess how this impacts genotyping with RNAseq.

Conclusion

In order to identify loci which matter for explaining quantitative trait variation or involved in adaptation to biotic or abiotic constraints, one needs to investigate a large number of individuals to reach a sufficient statistical power. But for a given amount of money to be spent in a sequencing experiment, there is a tradeoff between the sample size and the extent of the genome that can be examined [47]. Several methods have been proposed to reduce the complexity of the genome prior to sequencing enabling the multiplexing of individuals onto a sequencer lane. Here we have used RNAseq as a 'natural' alternative to reduce genome complexity prior to sequencing and have shown with several validations that it is efficient for the simultaneous discovery and typing of SNP. If all of these genome complexity reduction techniques have pros and cons [48–51], we believe that RNAseq has far been underexploited by comparison to the others and hope that our results will encourage its future use. One reason for the unpopularity of RNAseq for genotyping might be its cost which remains fairly expensive in comparison to GBS or RADseq, but one should also note that it also enables the access to the expression of genes in the tissue sampled which together with the SNPs generated can be used to detect eQTLs or ASE [13, 52].

Additional files

Additional file 1: Position of SNPs on *Populus trichocarpa* v3.0 genome version of the SNP identified by Marroni et al. [3, 4] on 5 genes. (XLSX 157 kb)

Additional file 2: Number of SNPs detected by different callers or combinations of callers and using different filters. **Table S1:** Total number of SNPs detected with four different callers and applying different filters.

Table S2: Comparison of number of SNPs detected with RNAseq data, identical positions with the SNP chip and genotyping accuracy using 7 calling modalities times 3 options for missing values. (XLSX 12 kb)

Additional file 3: Supplementary figures. **Figure S1:** Repetition quality control checked using genotyping from a previously available SNP chip [1]. **Figure S2:** Distribution of genotyping accuracy of RNAseq data computed from a comparison with genotyping from a previously available SNP array [1] for the 12 individuals used in the study. **Figure S3:** Variation of the total SNP number and identical positions found with the chip data using 7 calling modalities times 3 options for missing values. **Figure S4:** Positions of SNPs discovered and genotyped with RNAseq across 12 *Populus nigra* individuals and along two genes. **Figure S5:** Graphical representation of the enrichment in GO terms (biological process) for the genes covered by at least 5 SNPs. (PDF 434 kb)

Abbreviations

ASE: Allele-specific expression; BAM: Binary alignment map; BP: Base pair; BWA: Burrows-wheeler aligner; DNA: DeoxyriboNucleic acid; DNase: DeoxyriboNuclease; eQTL: expression quantitative trait locus; GATK: Genome analysis ToolKit; GBS: Genotyping by sequencing; GO: Gene ontology; GWAS: Genome-wide association study; LD: Linkage disequilibrium; MAF: Minor allele frequency; PCA: Principal component analysis; PCR: Polymerase chain reaction; PE: Paired-end; QTL: Quantitative trait loci; RAD: Restriction site associated DNA; RADseq: RAD sequencing; RNA: RiboNucleic acid; RNase: Ribonuclease; RNAseq: RNA sequencing; SAM: Sequence alignment map; SNP: Single nucleotide polymorphism; SR: Single-read; UTR: UnTranslated region; VCF: Variant call format

Acknowledgements

We thank Fabio Marroni for kindly providing original *P. nigra* NGS and reference sequences for "CAD4", "HCT1", "C3H3", "CCR7" and "4CL3".

Funding

This work was done within the SYBIOPOP project (ANR-13-JSV6-0001) funded by the French National Research Agency (ANR). The platform POPS benefits from the support of the LabEx Saclay Plant Sciences-SPS (ANR-10-LABX-0040-SPS).

Availability of data and materials

This RNAseq project has been submitted to the international repository Gene Expression Omnibus (GEO) from NCBI (accession number: GSE117346). All steps of the experiment, from growth conditions to bioinformatic analyses are detailed in CATdb [53] (Project: JC2013_SYBIOPOP_2014) according to the MINSEQE 'minimum information about a high-throughput sequencing experiment'. Raw sequences (fastq) have been deposited in the Sequence Read Archive (SRA) from NCBI (accession number: SRP154396). The final VCF is available in the INRA Dataverse repository. Information on the studied genotypes is available in the GnpIS Information System [54, 55].

Authors' contributions

VS designed the study; MCLD, SB, LST and JC carried out the experiments and the sequencing; OR, AC, SA, VB, VJ and VS contributed to data analysis; OR, AC, VJ and VS wrote the paper with input from all co-authors. All authors have read and approved the manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

¹BioForA, INRA, ONF, 45075 Orléans, France. ²Institute of Plant Sciences Paris-Saclay (IPS2), CNRS, INRA, Université Paris-Sud, Université Paris-Saclay, Université d'Evry, Université Paris-Diderot, Sorbonne Paris-Cité, 91405 Orsay, France. ³IRHS, INRA, Agrocampus-Ouest, Université d'Angers, SFR 4207 QUASAV, 49071 Beaucouzé, France.

Received: 17 July 2018 Accepted: 9 November 2018

Published online: 12 December 2018

References

- Faivre-Rampant P, Zaina G, Jorge V, Giacomello S, Segura V, Scalabrini S, Guérin V, De Paoli E, Aluome C, Viger M, Cattonaro F, Payne A, PaulStephenRaj P, Le Paslier MCC, Berard A, Allwright MRR, Villar M, Taylor G, Bastien C, Morgante M. New resources for genetic studies in *Populus nigra*: genome-wide SNP discovery and development of a 12k Infinium array. *Mol Ecol Resour.* 2016;16(4):1023–36. <https://doi.org/10.1111/1755-0998.12513>.
- Pinosio S, Giacomello S, Faivre-Rampant P, Taylor G, Jorge V, Le Paslier MC, Zaina G, Bastien C, Cattonaro F, Marroni F, Morgante M. Characterization of the Poplar Pan-Genome by Genome-Wide Identification of Structural Variation. *Mol Biol Evol.* 2016;33(10):2706–19. <https://doi.org/10.1093/molbev/msw161>.
- Marroni F, Pinosio S, Di Centa E, Jurman I, Boerjan W, Felice N, Cattonaro F, Morgante M. Large-scale detection of rare variants via pooled multiplexed next-generation sequencing: Towards next-generation Ecotilling. *Plant J.* 2011;67(4):736–45. <https://doi.org/10.1111/j.1365-313X.2011.04627.x>.
- Marroni F, Pinosio S, Zaina G, Fogolari F, Felice N, Cattonaro F, Morgante M. Nucleotide diversity and linkage disequilibrium in *Populus nigra* cinnamyl alcohol dehydrogenase (CAD4) gene. *Tree Genet Genomes.* 2011;7(5):1011–23. <https://doi.org/10.1007/s11295-011-0391-5>.
- Guerra FP, Wegrzyn JL, Sykes R, Davis MF, Stanton BJ, Neale DB. Association genetics of chemical wood properties in black poplar (*Populus nigra*). *New Phytol.* 2013;197(1):162–76. <https://doi.org/10.1111/nph.12003>.
- Allwright MR, Payne A, Emiliani G, Milner S, Viger M, Rouse F, Keurentjes JJB, Bérard A, Wildhagen H, Faivre-Rampant P, Polle A, Morgante M, Taylor G. Biomass traits and candidate genes for bioenergy revealed through association genetics in coppiced European *Populus nigra* (L.) *Biotechnol Biofuels.* 2016;9(1):195. <https://doi.org/10.1186/s13068-016-0603-1>.
- Slavov GT, Difazio SP, Martin J, Schackwitz W, Muchero W, Rodgers-Melnick E, Lipphardt MF, Pennacchio CP, Hellsten U, Pennacchio LA, Gunter LE, Ranjan P, Vining K, Pomraning KR, Wilhelm LJ, Pellegrini M, Mockler TC, Freitag M, Galdes A, El-Kassaby YA, Mansfield SD, Cronk QCB, Douglas CJ, Strauss SH, Rokhsar D, Tuskan GA. Genome resequencing reveals multiscale geographic structure and extensive linkage disequilibrium in the forest tree *Populus trichocarpa*. *New Phytol.* 2012;196(3):713–25. <https://doi.org/10.1111/j.1469-8137.2012.04258.x>.
- Galdes A, DiFazio SP, Slavov GT, Ranjan P, Muchero W, Hannemann J, Gunter LE, Wymore AM, Grassa CJ, Farzaneh N, Porth I, McKown AD, Skyba O, Li E, Fujita M, Klápště J, Martin J, Schackwitz W, Pennacchio C, Rokhsar D, Friedmann MC, Wasteneys GO, Guy RD, El-Kassaby YA, Mansfield SD, Cronk QCB, Ehlting J, Douglas CJ, Tuskan GA. A 34K SNP genotyping array for *Populus trichocarpa*: Design, application to the study of natural populations and transferability to other *Populus* species. *Mol Ecol Resour.* 2013;13(2):306–23. <https://doi.org/10.1111/1755-0998.12056>.
- Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE.* 2011;6(5). <https://doi.org/10.1371/journal.pone.0019379>. NIHMS150003.
- Miller MR, Dunham JP, Amores A, Cresko WA, Johnson EA. Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Res.* 2007;17(2):240–8. <https://doi.org/10.1101/gr.5681207>.
- Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, Shaffer T, Wong M, Bhattacharjee A, Eichler EE, Bamshad M, Nickerson DA, Shendure J. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature.* 2009;461(7261):272–6. <https://doi.org/10.1038/nature08250>.
- Fahrenkrog AM, Neves LG, Resende MFR, Vazquez AI, de los Campos G, Dervinis C, Sykes R, Davis M, Davenport R, Barbazuk WB, Kirst M. Genome-wide association study reveals putative regulators of bioenergy traits in *Populus deltoides*. *New Phytol.* 2016;213(2):799–811. <https://doi.org/10.1111/nph.14154>. arXiv:1011.1669v3.
- De Wit P, Pespeni MH, Palumbi SR. SNP genotyping and population genomics from expressed sequences - Current advances and future possibilities. *Mol Ecol.* 2015;24(10):2310–23. <https://doi.org/10.1111/mec.13165>.
- Galdes A, Pang J, Thiessen N, Cezard T, Moore R, Zhao Y, Tam A, Wang S, Friedmann M, Birol I, Jones SJM, Cronk QCB, Douglas CJ. SNP discovery in black cottonwood (*Populus trichocarpa*) by population transcriptome resequencing. *Mol Ecol Resour.* 2011;11(SUPPL. 1):81–92. <https://doi.org/10.1111/j.1755-0998.2010.02960.x>.
- McKown AD, Klápště J, Guy RD, Galdes A, Porth I, Hannemann J, Friedmann M, Muchero W, Tuskan GA, Ehlting J, Cronk QCB, El-Kassaby YA, Mansfield SD, Douglas CJ. Genome-wide association implicates numerous genes underlying ecological trait variation in natural populations of *Populus trichocarpa*. *New Phytol.* 2014;203(2):535–53. <https://doi.org/10.1111/nph.12815>.
- Porth I, Klápště J, Skyba O, Hannemann J, McKown AD, Guy RD, Difazio SP, Muchero W, Ranjan P, Tuskan GA, Friedmann MC, Ehlting J, Cronk QCB, El-Kassaby YA, Douglas CJ, Mansfield SD. Genome-wide association mapping for wood characteristics in *Populus* identifies an array of candidate single nucleotide polymorphisms. *New Phytol.* 2013;200(3):710–26. <https://doi.org/10.1111/nph.12422>.
- Konczal M, Koteja P, Orłowska-Feuer P, Radwan J, Sadowska ET, Babik W. Genomic Response to Selection for Predatory Behavior in a Mammalian Model of Adaptive Radiation. *Mol Biol Evol.* 2016;33(9):2429–40. <https://doi.org/10.1093/molbev/msw121>.
- Nürnberg B, Lohse K, Fijarczyk A, Szymura JM, Blaxter ML. Para-allopatry in hybridizing fire-bellied toads (*Bombina orientalis* and *B. variegata*): Inference from transcriptome-wide coalescence analyses. *Evolution.* 2016;70(8):1803–18. <https://doi.org/10.1111/evo.12978>.
- Summers CF, Gulliford CM, Carlson CH, Lillis JA, Carlson MO, Cadle-Davidson L, Gent DH, Smart CD. Identification of genetic variation between obligate plant pathogens *Pseudoperonospora cubensis* and *P. humuli* using RNA sequencing and genotyping-by-sequencing. *PLoS ONE.* 2015;10(11):0143665. <https://doi.org/10.1371/journal.pone.0143665>.
- Berthouly-Salazar C, Thuillet AC, Rhoné B, Mariac C, Ousseini IS, Couderc M, Tenaillon M, Vigouroux Y. Genome scan reveals selection acting on genes linked to stress response in wild pearl millet. *Mol Ecol.* 2016;25(21):5500–12. <https://doi.org/10.1111/mec.13859>.
- Lu X, Kracher B, Saur IML, Bauer S, Ellwood SR, Wise R, Yaeno T, Maekawa T, Schulze-Lefert P. Allelic barley MLA immune receptors recognize sequence-unrelated avirulence effectors of the powdery mildew pathogen. *Proc Natl Acad Sci.* 2016;113(42):6486–95. <https://doi.org/10.1073/pnas.1612947113>.
- Guét J, Fabbri F, Fichot R, Sabatti M, Bastien C, Brignolas F. Genetic variation for leaf morphology, leaf structure and leaf carbon isotope discrimination in European populations of black poplar (*Populus nigra* L.). *Tree Physiol.* 2015;35(8):850–63. <https://doi.org/10.1093/treephys/tpv056>.
- FastQC: A quality control tool for high throughput sequence data. 2010. <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>. Accessed 22 Nov 2018.
- Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet journal.* 2011;17(1):10. <https://doi.org/10.14806/ej.17.1.200>.
- FASTX toolkit. 2014. http://hannonlab.cshl.edu/fastx_toolkit/. Accessed 22 Nov 2018.
- Tuskan GA, Difazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A, Schein J, Sterck L, Aerts A, Bhalerao RR, Bhalerao RP, Blaudez D, Boerjan W, Brun A, Brunner A, Busov V, Campbell M, Carlson J, Chalot M, Chapman J, Chen G-L,

Phenomic Selection Is a Low-Cost and High-Throughput Method Based on Indirect Predictions: Proof of Concept on Wheat and Poplar

Renaud Rincant,* Jean-Paul Charpentier,^{†*} Patricia Faivre-Rampant,[§] Etienne Paux,*

Jacques Le Gouis,* Catherine Bastien,[†] and Vincent Segura^{†,1}

*GDEC, INRA, UCA, 63000 Clermont-Ferrand, France, [†]BioForA, INRA, ONF, 45075 Orléans, France, [‡]GenoBois analytical platform, INRA, 45075 Orléans, France, and [§]EPGV, INRA, CEA-IG/CNG, 91057 Evry, France

ORCID IDs: 0000-0003-0885-0969 (R.R.); 0000-0002-6029-0498 (J.-P.C.); 0000-0002-3094-7129 (E.P.); 0000-0001-5726-4902 (J.L.G.); 0000-0002-9391-6637 (C.B.); 0000-0003-1860-2256 (V.S.)

ABSTRACT Genomic selection - the prediction of breeding values using DNA polymorphisms - is a disruptive method that has widely been adopted by animal and plant breeders to increase productivity. It was recently shown that other sources of molecular variations such as those resulting from transcripts or metabolites could be used to accurately predict complex traits. These endophenotypes have the advantage of capturing the expressed genotypes and consequently the complex regulatory networks that occur in the different layers between the genome and the phenotype. However, obtaining such omics data at very large scales, such as those typically experienced in breeding, remains challenging. As an alternative, we proposed using near-infrared spectroscopy (NIRS) as a high-throughput, low cost and non-destructive tool to indirectly capture endophenotypic variants and compute relationship matrices for predicting complex traits, and coined this new approach "phenomic selection" (PS). We tested PS on two species of economic interest (*Triticum aestivum* L. and *Populus nigra* L.) using NIRS on various tissues (grains, leaves, wood). We showed that one could reach predictions as accurate as with molecular markers, for developmental, tolerance and productivity traits, even in environments radically different from the one in which NIRS were collected. Our work constitutes a proof of concept and provides new perspectives for the breeding community, as PS is theoretically applicable to any organism at low cost and does not require any molecular information.

KEYWORDS

Poplar
Wheat
breeding
endophenotypes
Near InfraRed
Spectroscopy
(NIRS)
Genomic
Prediction
GenPred
Shared Data
Resources

To meet the world's current and future challenges, especially in terms of food and energy supplies, there is a great need to develop efficient crop varieties, livestock breeds or forest materials through breeding. Until recently, the selection of promising individuals in animal and plant breeding was mostly based on their phenotypic records. This approach was a strong limit to genetic progress as the high costs of phenotyping

strongly constrain the number of candidates that can be evaluated, especially when there are interactions between individuals and environments that necessitate the evaluation of selection candidates in various environments. Another strong constraint - typical in perennial crops, trees or animals - is that it can sometimes take several years to evaluate phenotypes, which increases the duration of selection cycles. These limitations are some of the main reasons why genomic selection (GS) has become so popular in the last two decades. Its principle is based on a combination of phenotypic records and genome-wide molecular markers to train a prediction model that can in turn be used to predict the performances of - potentially unphenotyped - individuals (Meuwissen *et al.* 2001). We can thus select more individuals faster, which increases genetic gain. The development of high-throughput genotyping tools at decreasing costs has made GS possible for many animal and plant species. It can be used both in pre-breeding to screen diversity material (Crossa *et al.* 2016; Yu *et al.* 2016; Gorjanc *et al.* 2016) and in breeding to make the schemes more efficient (Heffner *et al.* 2010;

Copyright © 2018 Rincant *et al.*

doi: <https://doi.org/10.1534/g3.118.200760>

Manuscript received September 26, 2018; accepted for publication October 20, 2018; published Early Online October 29, 2018.

This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Supplemental material available at Figshare: <https://doi.org/10.25387/g3.7243256>.

¹Corresponding author: BioForA, INRA, ONF, 45075 Orléans, France. E-mail: vincent.segura@inra.fr

Meuwissen *et al.* 2013; Gaynor *et al.* 2017). However, a great number of species are still orphans of any genotyping tool, and for many others, genotyping costs remain a limit to the implementation of GS in pre-breeding and breeding. In addition, genotyping thousands to millions of individuals (potentially each year) is a challenge that consequently remains inaccessible for most species, even if the cost efficiency has improved thanks to low coverage genotyping by sequencing (Elshire *et al.* 2011) and imputation (Gorjanc *et al.* 2017a,b).

One of the reference GS models is the ridge regression BLUP (RR-BLUP, (Whittaker 2000; Meuwissen *et al.* 2001)) in which a penalized regression is made on all markers simultaneously. This model assumes that the genes affecting the trait of interest are spread across the whole genome and that all of these genes have small effects. Despite its simplicity, this model has been proven one of the most effective in many situations, except for when major genes contribute to trait architecture. Interestingly, this model is equivalent to the genomic BLUP model (G-BLUP, (Habier *et al.* 2007; Goddard 2009; Hayes *et al.* 2009; Zhong *et al.* 2009)) in which markers are used to estimate a realized genomic relationship matrix between individuals, also called kinship. This framework means that we can compress genome-wide information from numerous molecular markers into summary statistics (kinship coefficients between individuals) without diminishing prediction accuracy. Considering this fact, we should ask the question: are there more efficient alternatives than genotyping to estimate the kinship matrix? In the last years, it was proposed to use endophenotypes (Mackay *et al.* 2009) such as transcripts (Fu *et al.* 2012; Guo *et al.* 2016; Zenke-Philippi *et al.* 2017; Westhues *et al.* 2017), small RNAs (Seifert *et al.* 2018) or metabolites (Riedelsheimer *et al.* 2012; Feher *et al.* 2014; Ward *et al.* 2015; Fernandez *et al.* 2016; Xu *et al.* 2016; Guo *et al.* 2016; Schrag *et al.* 2018) as regressors or to estimate kinship. These endophenotypes correspond to different molecular layers between the genome and the phenotype, which permits the integration of interactions and regulatory networks. These kinds of variables have proven to be efficient to predict integrative traits using the same statistical models as those classically used in GS. These regressors have the advantage of capturing expressed genotypes, but they remain too expensive to be routinely applied on the large scales typically dealt with by breeders. It is interesting to note that even with a small portion of the transcripts or metabolites sampled on a single tissue in a single environment and sometimes at very early stages, it was possible to compute kinship matrices allowing to reach predictive abilities similar to those obtained with molecular markers (Riedelsheimer *et al.* 2012; Xu *et al.* 2016). One could thus consider the possibility of using cheaper and easier techniques to capture endophenotypic variations.

Near-infrared spectroscopy (NIRS) is a high-throughput, non-destructive and low-cost method routinely used to estimate reflectance of a sample for numerous wavelengths. This reflectance is mainly related to the presence of chemical bonds in the analyzed tissue and as a result is expected to be related to endophenotypes. We suppose that the reflectance at each of the numerous wavelengths can be considered as an integration of numerous endophenotypic variations. We thus propose to evaluate the efficiency of NIRS to make predictions with G-BLUP (or equivalently, RR-BLUP) using these traits instead of molecular markers. Numerous studies have demonstrated the usefulness of NIRS for barcoding samples and discriminating species or varieties (Bertrand *et al.* 1985; Adedipe *et al.* 2008; Espinoza *et al.* 2012; Fischnaller *et al.* 2012; Abasolo *et al.* 2013; O'Reilly-Wapstra *et al.* 2013; Meder *et al.* 2014; Lang *et al.* 2017) and have thus suggested that NIRS could be considered as a genetic marker (Cruickshank and Munck 2011).

Moreover, some studies have shown that NIRS can capture some genetic variability by estimating the heritability of absorbances along the spectrum and even mapping corresponding quantitative trait loci (QTL, (Posada *et al.* 2009; Diepeveen *et al.* 2012; O'Reilly-Wapstra *et al.* 2013; Hein and Chaix 2014)). However, to the best of our knowledge, no studies have proposed using NIRS to perform "phenomic selection" (PS), which we define as the use of high-throughput phenotyping to obtain numerous variables which can be used as regressors or to estimate kinship in the statistical models classically used in GS. We emphasize that the concept of phenomic selection is radically different from the classical use of NIRS prediction. In the classical methodology, NIRS is collected on a sample to make prediction on that particular sample for traits of various complexity (from chemical composition (Foley *et al.* 1998) to yield (Ferrio *et al.* 2005; Cabrera-Bosquet *et al.* 2012; Weber *et al.* 2012; Aguate *et al.* 2017)) using a formula that has previously been calibrated. On the other hand in PS, NIR reflectances are considered in the same way as genomic or endophenotypic regressors, at the genotypic level rather than the individual level, which allows making predictions in any environment without having any environment specific NIRS. In PS we suppose that once NIR reflectances are analyzed in one experiment (collections of seed, a nursery, a trial or a controlled experiment) they could be used as regressors or to estimate a kinship matrix to make predictions in any other experiment, as long as relevant phenotypic data are available to calibrate the statistical model, like in GS with molecular markers.

There are several advantages to this approach. One can obtain NIRS for any plant or animal species at a lower cost than genotyping and potentially without particular treatment of the samples prior to the analysis such as DNA or RNA extraction. One can also obtain NIRS directly in the field thanks to portable devices (Ecartot *et al.* 2013; Teixeira Dos Santos *et al.* 2013) or autonomous high-throughput vectors, such as phénomobiles (Madec *et al.* 2017) that generate hyper-spectral images (Diago *et al.* 2013; Peerbhay *et al.* 2013). NIRS can even be obtained non destructively on seeds before sowing. As a result, prediction-based selection would be possible for any species and at a low enough cost to make it interesting to implement, even if its results are less accurate than those of GS. As a proof of concept of PS, we report an evaluation of the usefulness of NIRS for predicting quantitative traits of economic interest within two different species, a cereal (winter wheat) and a tree (poplar) using various tissues (grains, leaves, wood) and under different environments, and compare the results to those of a GS prediction based on several thousand SNPs.

MATERIALS AND METHODS

Data

Genetic material and experimental designs: *Wheat* The panel was composed of 228 European elite varieties of winter wheat released between 1977 and 2012, 89% of which have been released since 2000. 72.8% of these varieties are in the panel introduced in Ly *et al.* (2018). The full panel was sown in one trial in Clermont-Ferrand (France) in 2015/2016. This trial was an augmented design with two treatments: one drought treatment under rain-out shelters (DRY), and one irrigated treatment (IRR) next to it. There was a difference of 223 mm in water supply (rainfall and irrigation) between the two treatments at the end of the experiment. For both treatments, the panel was divided into eight blocks of earliness with one replicate within the same block for 64 varieties and no replicates for the other 164, except for four checks, which were replicated three times in each block. Phenotypes and NIRS were collected in these two reference environments. A subset of 161 varieties were sown and phenotyped in six

independent environments located in Estrées-Mons (France, 2011/2012 and 2012/2013) and Clermont-Ferrand (France, 2012/2013) with two treatments corresponding to two levels of nitrogen input (intermediate and high). This subpanel was divided into six groups of earliness and each group was repeated in two blocks. Four checks were present in each block.

Poplar The population was an association population comprising 1,160 cloned genotypes representative of the natural range of the species in Western Europe and previously described (Guet *et al.* 2015; Faivre-Rampant *et al.* 2016; Gebreselassie *et al.* 2017). Clonally replicated trials of subsets of this association population were established in 2008 at two contrasting sites in central France (Orléans, ORL) and Northern Italy (Savigliano, SAV), with 1,098 and 815 genotypes at ORL and SAV respectively. At each site, a randomized complete block design was used with a single tree per block and six replicates per genotype. Growth data collected in each design clearly indicated that the Italian site was more favorable than the French site (Guet *et al.* 2015; Gebreselassie *et al.* 2017).

NIRS data: Wheat NIRS data were obtained on flag leaves and harvested grains from the two treatments of the drought trial in Clermont-Ferrand (France) in 2015/2016. For each variety in each treatment, twenty flag leaves were sampled on one plot at 200 degree days after flowering. The samples were oven dried at 60° for 48 h. Leaves were milled (Falaise miller, SARL Falaise, France), and the powder was analyzed with a FOSS NIRS 6500 (FOSS NIRSystems, Silver Spring, MD) and its corresponding softwares (ISIScan™ and WINisi™ 4.20). For each variety in each treatment, 200 g of grains harvested at one plot were analyzed with a FOSS NIRS XDS (FOSS NIRSystems, Silver Spring, MD) and its corresponding softwares (ISIScan™ and WINisi™ 4.20). For leaf powder and grain, absorbance was measured from 400 to 2500 nm with a step of 2 nm. 5 varieties were removed from the dataset because their leaf absorbance was abnormal because of a technical problem, resulting in a final panel of 223 varieties. The resulting spectra were loaded into R software (R Core Team 2018) to be pretreated using custom R code. They were normalized (centered and scaled) and their first derivative was computed using a Savitzky-Golay filter (Savitzky and Golay 1964) with a window size of 37 data points (74 nm) implemented in the R package signal (Signal Developers 2014). In the end, each variety in each treatment was characterized by a transformed spectrum (first derivative of the normalized spectrum) of flag leaf powder and a transformed spectrum of grains.

Poplar NIRS was carried out on wood from stem sections collected at 1 m above ground on 2-year-old trees for 1,081 genotypes in three blocks at Orléans (total of 2,860 samples) and 792 genotypes in three blocks at Savigliano (total of 2,254 samples). After harvest, the wood samples were oven dried at 30° for several days, cut into small pieces with a big cutter and milled using a Retsch SM2000 cutting mill (Retsch, Haan, Germany) to pass through a 1-mm sieve. The wood samples were not debarked prior to milling. After stabilization, wood powders were placed into quartz cups for NIR collection with a Spectrum 400 Fourier-transformed spectrometer (Perkin Elmer, Waltham, MA, USA) and its corresponding software (Spectrum™ 6.3.5). For each sample, the measurement consisted of an average of 64 scans done while rotating the cups over the 10,000 cm⁻¹ - 4,000 cm⁻¹ range with a resolution of 8 cm⁻¹ and a zero-filling factor of 4, resulting in absorbance data every 2 cm⁻¹. The resulting spectra were loaded into R software (R Core Team 2018) to be processed using custom R code. They were first restricted to the 8000 cm⁻¹ - 4000 cm⁻¹ range because the most distant part of the spectra (8000 cm⁻¹ - 10,000 cm⁻¹) appeared to be quite noisy. Then, the restricted spectra were normalized (centered and scaled), and their first derivative was computed using a Savitzky-Golay filter (Savitzky and Golay 1964) with a window size of 37 data points

(74 cm⁻¹) implemented in the R package signal (Signal Developers 2014). Finally, these normalized and derived spectra were averaged by genotype at each site.

SNP data: Wheat The 228 wheat varieties were genotyped with the TaBW280K high-throughput genotyping array described in Rimbart *et al.* (2018). This array was designed to cover both genic and intergenic regions of the three subgenomes. Markers with a minor allele frequency below 1%, or with a heterozygosity or missing rate above 5% were removed. Groups of identical markers were identified, and for each of them only one marker was kept. Eventually, we obtained 84,259 SNPs, either polymorphic high resolution or off-target variants, with an average missing data rate of 0.83%. Missing values were imputed as the marker frequency.

Poplar The poplar association population was genotyped with an Illumina Infinium BeadChip array (Faivre-Rampant *et al.* 2016) yielding 7,918 SNPs for 858 genotypes. Missing values were rare (0.35%) and they were imputed with FImpute (Sargolzaei *et al.* 2014). The data were restricted to the subset of 562 genotypes with SNP data and NIRS data at both sites. Within this set, SNPs with a minor allele frequency below 1% were discarded, yielding a final SNP dataset of 7,808 SNPs.

Phenotypic data: Wheat The 228 wheat varieties were phenotyped for heading date (HD) and grain yield (GY) at the two environments in which the NIRS analysis was conducted (drought experiment in Clermont-Ferrand 2015/2016). The subpanel of 161 varieties was phenotyped for the same traits in six independent environments. In each environment, the phenotypic data were adjusted for micro-environmental effects using the random effect block and when necessary by modeling spatial trends using two-dimensional penalized spline (P-spline) models as implemented in the R package SpATS (Rodríguez-Álvarez *et al.* 2017). Broad-sense heritabilities were computed following Oakey *et al.* (2007).

Poplar The poplar association population was evaluated at each of the two sites for the following traits on up to six replicates by genotype: height at 2 years at Orléans (HT-ORL), circumference at 1 m above ground at 2 years at both sites (CIRC-ORL and CIRC-SAV), bud flush at both sites (BF-ORL and BF-SAV) and bud set at both sites (BS-ORL and BS-SAV) as discrete scores for a given day of the year (see Dillen *et al.* (2009) and Rohde *et al.* (2011) for details on the scales used) and resistance to rust at Orléans (RUST-ORL) as a discrete score of susceptibility on the most affected leaf of the tree and on a 1 to 8 scale. Within each site, the phenotypic data were adjusted for micro-environmental effects using random effect block and/or spatial position when needed following a visual inspection of spatial effects with a variogram as implemented in the R package breedR (Muñoz and Sanchez 2017). Finally, the adjusted phenotypes were restricted to the subset of 562 genotypes with SNP and NIRS data for computing an averaged genotypic value for each trait by genotype within each site for further analyses.

Genomic heritability and partition of variance along spectra

The estimation of genomic heritability was based on the following bivariate statistical model across environments:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}, \quad (1)$$

where y_1 , y_2 are the phenotypic values (absorbance for a given wavelength) in each environment, $\boldsymbol{\beta}$ is a vector of fixed environment effect, \mathbf{u} is a vector of random polygenic effect with

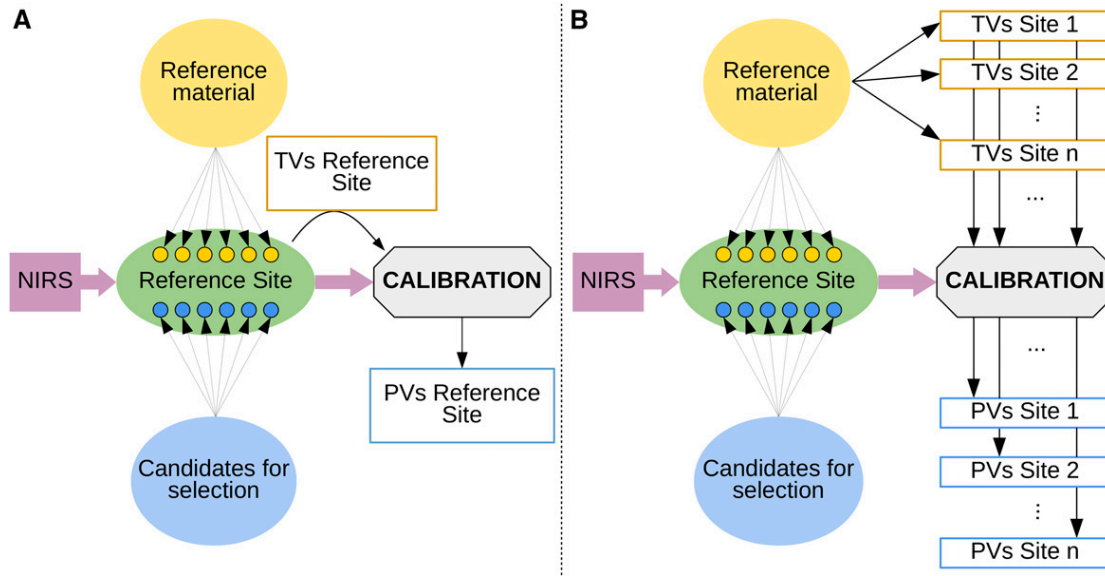


Figure 1 Schematic representation of the concept of phenomic selection, including the two scenarios tested in the present work: (a) S1, where the calibration model is trained with true values (TVs) and NIRS data collected at the same - reference - site and (b) S2, where the calibration model is trained with NIRS data collected at the reference site and TVs from other(s) environment(s). In both scenarios, the outcome of the prediction consists of predicted values (PVs).

$var(\mathbf{u}) = \begin{bmatrix} \sigma_{u_1}^2 & \sigma_{u_{12}} \\ \sigma_{u_{21}} & \sigma_{u_2}^2 \end{bmatrix} \otimes \mathbf{K}$, \mathbf{K} being the scaled realized relationship matrix (see below), \mathbf{e} is a vector of independent and normally distributed residuals with $var(\mathbf{e}) = \begin{bmatrix} \sigma_{e_1}^2 & 0 \\ 0 & \sigma_{e_2}^2 \end{bmatrix} \otimes \mathbf{I}$, \mathbf{X} and \mathbf{Z} are design matrices relating observations to the effects. SNPs were used to estimate the genomic relationship matrix (\mathbf{A}) between individuals, following the formula of VanRaden (VanRaden 2008)

$$\mathbf{A} = \frac{1}{L} \sum_{l=1}^L \frac{(G_{i,l} - p_l)(G_{j,l} - p_l)}{\sigma^2}, \quad (2)$$

where $G_{i,l}$ and $G_{j,l}$ are the genotypes of individuals i and j at marker l ($G_{i,l} = 0$ or 1 for homozygotes, 0.5 for heterozygotes), p_l is the frequency of the allele coded 1 for the marker l , and σ^2 is the average empirical marker genotype variance. \mathbf{K} was obtained by scaling \mathbf{A} to have a sample variance of 1 (Kang *et al.* 2010; Forni *et al.* 2011). Genomic heritability was estimated for each wavelength within each environment (m) as follows:

$$h_m^2 = \frac{\hat{\sigma}_{u_m}^2}{\hat{\sigma}_{u_m}^2 + \hat{\sigma}_{e_m}^2}, \quad (3)$$

with $\hat{\sigma}_{u_m}^2$ and $\hat{\sigma}_{e_m}^2$ the REML estimates of $\sigma_{u_m}^2$ and $\sigma_{e_m}^2$, obtained with the Newton-Raphson algorithm implemented in the R package sommer (Covarrubias-Pazarán 2016). Following Yamada *et al.* (1988), the variance/covariance estimates from the previously defined bivariate mixed-model were used to compute estimates of genetic ($\hat{\sigma}_G^2$), genetic by environment ($\hat{\sigma}_{G \times E}^2$) and residual ($\hat{\sigma}_\epsilon^2$) variances across sites as follows: $\hat{\sigma}_G^2 = \hat{\sigma}_{u_{12}}$, $\hat{\sigma}_{G \times E}^2 = \frac{1}{2}(\hat{\sigma}_{u_1}^2 + \hat{\sigma}_{u_2}^2) - \hat{\sigma}_{u_{12}}$, and $\hat{\sigma}_\epsilon^2 = \frac{1}{2}(\hat{\sigma}_{e_1}^2 + \hat{\sigma}_{e_2}^2)$.

Association mapping on NIRS absorbance

Association mapping was carried out along spectra considering the absorbance at a given wavelength as a bivariate trait (corresponding to the two environments) and using previous estimates of genetic and

residual variances (EMMAX philosophy as previously proposed in the multi-trait mixed-model approach (Korte *et al.* 2012)).

Phenotype prediction with genomic and phenomic information

The efficiency of genomic and phenomic predictions was evaluated by cross-validations in two types of scenarios (Figure 1). In scenario S1, NIRS analysis and cross-validation were applied to the same environment (Figure 1 a). In scenario S2, cross-validation was applied to independent environments: the environment(s) in which NIRS was collected and the environment in which the cross-validation was applied (calibration and prediction) were different (Figure 1 b). In S1, the objective was to limit expensive or labor-demanding phenotyping to a calibration set of reduced size and to predict the remaining individuals using NIRS. In scenario S2, one experiment (or a nursery) was dedicated to collecting the NIRS of the calibration set and the predicted set, and a multi-environment trial was dedicated to phenotyping the calibration set. The main difference between S1 and S2 was that in S1, we can expect NIRS to be more related to the phenotypic data than in S2.

For both scenarios, 5- and 8-fold cross-validation procedures repeated 20 times were used for poplar and wheat, respectively. A larger fold-number was considered for wheat in comparison to poplar because the sample size in the wheat dataset ($n = 223$ in the panel, and $n = 161$ in the subpanel) was lower than the sample size in the poplar dataset ($n = 562$). We would like to emphasize that in both scenarios, the predicted set was only characterized by genotypic and NIRS (obtained in the reference site) data. In particular in scenario S2, the only information available from the predicted environment is the phenotypic data of the calibration set (Figure S1). Predictive ability was computed as the Pearson correlation between the predictions and adjusted means. For genomic predictions, we tested two complementary reference models: G-BLUP and Bayesian LASSO (Park and Casella 2008; de los Campos *et al.* 2009). The underlying assumptions of these two models are that the SNP effects are normally distributed for G-BLUP, whereas Bayesian LASSO allows for departure from normality (*i.e.*, SNPs with bigger

effects). G-BLUP and Bayesian LASSO were run with the R packages rrBLUP (Endelman 2011) and BGLR (de los Campos *et al.* 2013), respectively. For Bayesian LASSO, the chain was composed of 30,000 iterations with a burn-in of 5,000 iterations, and the hyperparameter λ was chosen as recommended in de los Campos *et al.* (2013). For phenomic predictions, we used RR-BLUP but considered NIRS data instead of molecular markers. Prior to the analysis, the pretreated NIRS matrices were centered and scaled for each wavelength. For wheat, we also tested to include the spectrum of the two tissues (or of the same tissue but collected in the two environments) in the statistical model. For this we simply joined the two matrices of spectrum into one single matrix.

Expected genetic gain with genomic and phenomic selection in a simple example

We ran simulations to illustrate the expected genetic gain with GS and PS that would be achieved in one cycle of selection for various combinations of costs and reliabilities. Reliability was defined as the squared correlation between true breeding values (TBV) and the genomic or NIRS predicted values (PV).

We considered a situation in which a given budget (200,000 €) was available to predict the performances of selection candidates with NIRS or genotyping. Depending on the costs of the methods (DNA extraction and genotyping for GS or tissue sampling and NIRS acquisition for PS), we computed the number of selection candidates (N) that could be analyzed. The TBV and genomic or NIRS PV of these N individuals were then sampled from a multivariate normal distribution with means equal to 0, variances equal to 1 and covariance equal to the square root of reliability (R package *mvtnorm* (Genz *et al.* 2018)). The expected genetic gain was then computed as the difference between the average TBV of the 400 individuals having the best PV and the average TBV of the population (equal to 0). We selected 400 individuals because for many species, it is feasible to apply heavier phenotyping (multi-environment trials) on a few hundred individuals. We considered two situations; in the first situation, the expected genetic gain of GS and PS was computed for various genotyping and NIRS costs with a reliability set to 0.4. In the second situation, the reliability of GS and PS varied between 0.3 and 0.6, and genotyping and NIRS costs were set to 50 € and 4 €, respectively. For each combination of parameters (reliabilities and costs of GS and PS), the simulation procedure was repeated 1000 times to obtain stable results. Because genotyping and NIRS costs are highly dependent on the species and the number of samples analyzed, we let the genotyping costs (DNA extraction and genotyping itself) vary between 25 € and 100 € and the NIRS costs (sample treatment and NIRS analysis itself) vary between 1 € and 8 € in the first situation.

To provide concrete examples, we applied this simulation process with the reliabilities and costs that we experienced for wheat and poplar. GS costs were between 35 € and 50 € per individual for wheat and poplar, respectively, and PS costs were between 3 € and 2.5 € per individual for wheat and poplar, respectively. Reliabilities were estimated as the square of predictive abilities estimated by cross-validation divided by the heritability of the adjusted means. For each combination of trait, scenario, and NIRS data considered (tissue, environment), the increase in expected genetic gain using PS instead of GS was computed with the best performing GS model as a reference.

Data Availability

The datasets generated during and/or analyzed during the current study are available in the INRA Dataverse repository (<https://data.inra.fr/>). They can be accessed with the following link <http://dx.doi.org/10.15454/MB4G3T>. R functions used for comparing the predictive

ability of SNP and NIRS through cross-validations have been deposited on github (<https://github.com/viseadura/PS>). Supplemental material available at Figshare: <https://doi.org/10.25387/g3.7243256>.

RESULTS

Genomic heritability and partition of variance along NIRS

We first sought to characterize the ability of NIRS to capture genetic variability by estimating genomic heritability and partitioning the variance into genetic (G), genetic by environment ($G \times E$) and residual variances (ϵ) along the NIR spectrum collected on a panel of winter wheat (leaves and grains) and a population of black poplar (wood) grown in two contrasting environments. For both species and tissues, genomic heritability was highly variable along the spectrum with peaks above 60%, showing the existence of strong polygenic signals for some wavelengths (Figure 2, Figure S2, and File S1). For a given species, the proportion of $G \times E$ variance was on average across all the wavelengths equal to 8% (poplar), 10% (wheat leaves) or 16% (wheat grains) (Figure 2 and File S1). It is interesting to note that for at least half of the wavelengths, the cumulative proportion of G and $G \times E$ variances was above 15%, showing that the NIR signal was often partially related to genetics. The kind of tissue analyzed by NIRS seemed to matter, as shown by the comparison of variance partition along spectra obtained on wheat leaves and grains. G and $G \times E$ variances were higher and more stable along the spectrum for grains than for leaves.

We ran association mapping along the NIR spectrum to identify wavelengths associated with major QTL (Figure S2). In poplar, the signal appeared to be mainly polygenic with very few QTL detected, and the largest SNP R^2 was below 0.025 for any wavelength. In contrast, in winter wheat, we detected numerous large-effect QTL. For some wavelengths, a single SNP could have an R^2 of 0.23 for leaves and of 0.11 for grains, and this SNP could be in spectrum regions of high or of low genomic heritability (Figure 2). This finding means that depending on the wavelength, NIRS could capture highly polygenic relationships (wavelengths with high genomic heritability) or could tag specific regions of the genome (major QTL). These two kinds of wavelengths can be useful for making predictions because they can potentially track the two main factors responsible for GS accuracy: relatedness and linkage disequilibrium.

Comparing predictive abilities obtained with markers and with NIRS

We estimated the efficiency of GS and PS to predict the performance of new individuals within a cross-validation framework. The performances of the individuals in the validation set were predicted with genotypic information in GS (G-BLUP and Bayesian LASSO models) and with NIRS only in PS (RR-BLUP model). We considered two scenarios: in S1, NIRS analysis and cross-validation were performed in the same environment (Figure 1 a), whereas in S2, the environments in which the cross-validation was applied were different from those in which NIRS was obtained (Figure 1 b). The broad-sense heritabilities of the adjusted means were above 0.8 for all traits in each environment (Table S1).

In wheat, the predictive abilities of PS were highly variable and appeared to be dependent on the predicted trait and on the environment and tissue in which NIRS was measured (Figure 3 a, b, c, d and Figure 4). While combining NIRS collected in different environments or different tissues increased the predictive ability, this increase did not occur systematically. One major result is that for both traits, NIRS could lead to better predictions than molecular markers, even in the six independent

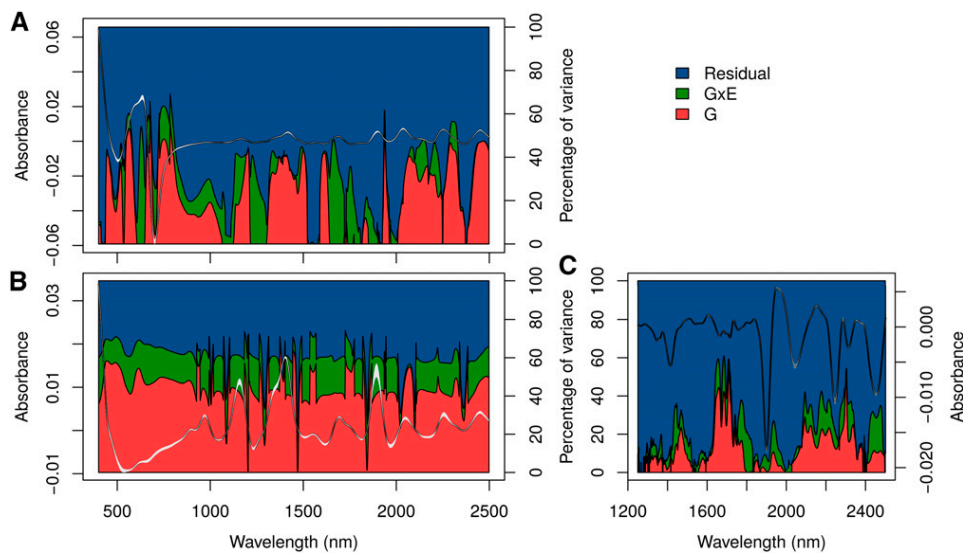


Figure 2 Proportion of genetic (red), genetic by environment (green), and residual (blue) variances along the NIR spectrum of (a) winter wheat leaves, (b) winter wheat grains and (c) poplar wood. NIRS was performed on plant material collected on genotypes grown under favorable and unfavorable environmental conditions. The median normalized and derived spectra, along with their first and third quartiles across the genotypes under study, are indicated in gray.

environments (Figure 3 c, d and Figure 4). The gain with NIRS in comparison to molecular markers in S2 was up to 34% and 22% for heading date and grain yield, respectively. In each S2 environment and for both traits, there was always a type of NIRS that performed better or as well as the best GS model (Figure 4). The gain was even stronger in S1: NIRS led to an increase in predictive ability of up to 53% and 117% for heading date and grain yield, respectively. In poplar, the predictive abilities with NIRS were always lower than those with SNP, except for growth traits under S1 (Figure 3 e). In the other cases, the predictive ability with NIRS varied depending on the trait and scenario considered, but they were always significantly greater than 0. In general, they were higher when the spectra were collected in the same environment (S1) than when spectra from another environment were used (S2), except for bud flush evaluated in one site and bud set evaluated in another site. Interestingly, irrespectively of the scenario, for some traits apparently unrelated to wood chemical properties, such as resistance to rust or bud set, NIRS predictive abilities were fairly high ranging between 0.34 and 0.53.

Expected genetic gain with genomic and phenomic selection in a simple example

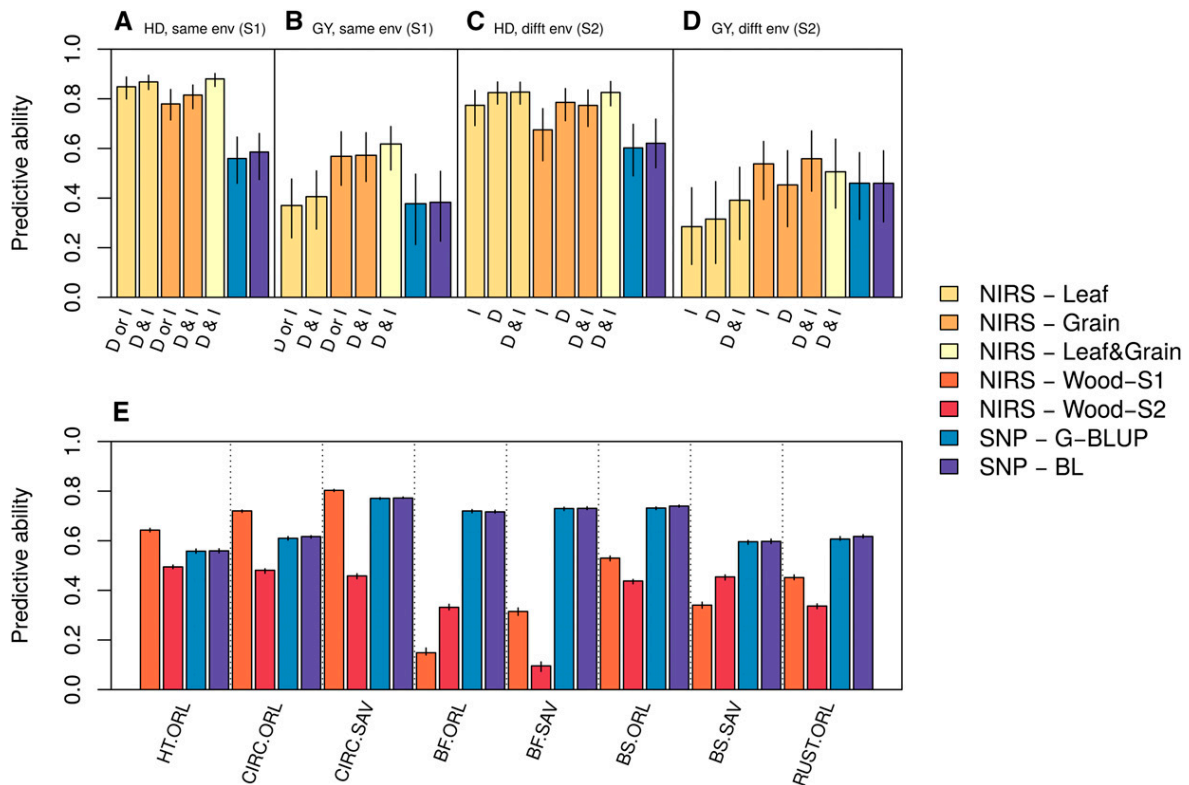
To further evaluate the potential of PS with respect to GS, the expected genetic gain with both approaches was compared in a simple scenario in which a budget of 200,000 € could be spent to genotype or analyze the NIRS of selection candidates. The difference in efficiency between GS and PS was highly dependent on the genotyping and NIRS costs and on the reliability of the two approaches (Figure 5). In the scenarios that we considered here, the expected gain of using PS instead of GS was between 11% and 127%. In extreme scenarios in which genotyping was cheap (25 €) and NIRS was expensive (8 €) or in which GS reliability (0.6) was much higher than PS reliability (0.3), PS was still better than GS. We applied the simulation process with the reliabilities and costs obtained in the wheat example (35 € for genotyping and DNA extraction and 3 € for sample treatment and NIRS acquisition). The increase of expected genetic gain with PS in comparison to GS was between +60% and +127% for heading date and between -10% and +222% for grain yield, depending on the tissue and environment used for NIRS acquisition and scenario considered (Table S2). In poplar, considering genotyping and NIRS acquisition costs of 50 € and 2.5 €, respectively, as well as the reliabilities estimated with cross-validation predictive

abilities, the expected gain in genetic progress varied depending on the trait and scenario considered (Table S3). It was mainly positive for growth traits (-2–93%), bud set (-6–25%) and rust resistance (-10–21%), whereas for bud flush, NIRS prediction did not seem to provide any advantage over regular SNP-based prediction.

DISCUSSION

In typical plant breeding programs, breeders have to select among thousands to millions of individuals. For most individuals, this selection is often based on a very small amount of phenotypic information because it is too expensive or simply impossible to make a precise phenotypic evaluation. It is also difficult and too expensive to genotype all individuals to apply GS, despite important economies of scales. Alternative approaches based on endophenotypes such as transcriptomes or metabolomes have been proposed to predict phenotypes (Fu *et al.* 2012; Riedelsheimer *et al.* 2012; Feher *et al.* 2014; Ward *et al.* 2015; Fernandez *et al.* 2016; Guo *et al.* 2016; Xu *et al.* 2016; Zenke-Philippi *et al.* 2017; Westhues *et al.* 2017; Seifert *et al.* 2018; Schrag *et al.* 2018), but their relatively low throughput and high costs are still likely to hamper their deployment at a large scale. To increase genetic progress in this context, we propose a new approach in which we use NIRS as high-throughput phenotypes to make predictions at low costs. The basic idea of this approach, which we call “phenomic selection” (PS), is that the absorbance of a sample in the near-infrared range is mainly related to its chemical composition, which depends itself on endophenotypes and genetics. Therefore, NIRS is supposed to capture at least part of the genetic variance, and as a result, one could use it to make predictions of traits unrelated to the analyzed tissue or in independent environments. The process of PS is similar to GS, but instead of reference material and selection candidates being genotyped, they are analyzed by NIRS.

We applied PS to the NIR spectrum of different tissues sampled on an association population of poplar and a panel of elite winter wheat. By estimating the extent of genetic variance along the NIR spectrum of poplar wood and winter wheat leaves and grains, we could show that most wavelengths displayed genetic variability (Figure 2). This result agrees with previous findings with eucalyptus wood (Hein and Chaix 2014), but whether this will still be true within pedigrees with a narrower genetic basis remains to be assessed. O’Reilly-Wapstra *et al.* (2013) have shown that NIR spectra collected on eucalyptus leaves



could differentiate full-sibs, even though the extent of genetic variation captured was lower than at the inter-specific level. Still these results suggest that NIRS could be valuable to capture some Mendelian sampling and that PS would work within pedigrees, but this hypothesis should clearly be tested in future work.

In the present work, the NIR spectra were specific to the environments in which they were obtained, but when they were analyzed jointly, we observed that G variance was larger than $G \times E$ variance for most wavelengths in both species. Posada *et al.* (2009) also reported a similar trend with coffee grains. This finding shows that even if the absorbances were partly environment specific, it should be possible to make predictions in independent environments. This result was further demonstrated by the good predictive abilities obtained with PS for most phenotypes in both species in scenario S2, *i.e.*, when the environment in which we trained the calibration model was different from the environment in which we collected NIRS. For both species, PS abilities were in the same range as GS abilities, sometimes performing better and sometimes performing worse than one another. For wheat, the results were very encouraging as we always found a situation (combination of environment and tissue analyzed) for which NIRS performed better than GS, even in six independent environments. More importantly, even when the correlation between the S1 and S2 environment was

as low as 0.16 for the predicted trait (Table S4, GY in Mon12N-), PS could produce better predictions than GS (Figure 4). In other words, a prediction model based on NIRS obtained in one specific environment could be used to make predictions in completely different environments. These promising results obtained in scenarios S1 and S2 open the way to important opportunities in the plant breeding community. As revealed by our theoretical computations (Figure 5), we expect PS to be able to generate large gains in genetic progress in comparison to GS, even in pessimistic scenarios. In the realistic scenarios that we experienced, the expected gain brought by using PS instead of GS could be up to 81% for wheat grain yield in scenario S2 (Table S2).

Nevertheless, these simulations have to be considered with caution, because of the strength of the underlying hypotheses. Our work has shown the interest of the proposed PS approach within a given generation that may clearly be applicable within plant breeding programs to assess the performance of the candidate for selection. But, more work is clearly needed to establish the proportion of the variance along NIRS (and for endophenotypes) that is heritable in the narrow-sense and thus transmitted to the next generations to be further used at different stages of the breeding programs or in different breeding contexts. Indeed, similarly to endophenotypes we expect NIRS to capture non-additive genetic effects which may overestimate the expected genetic progress

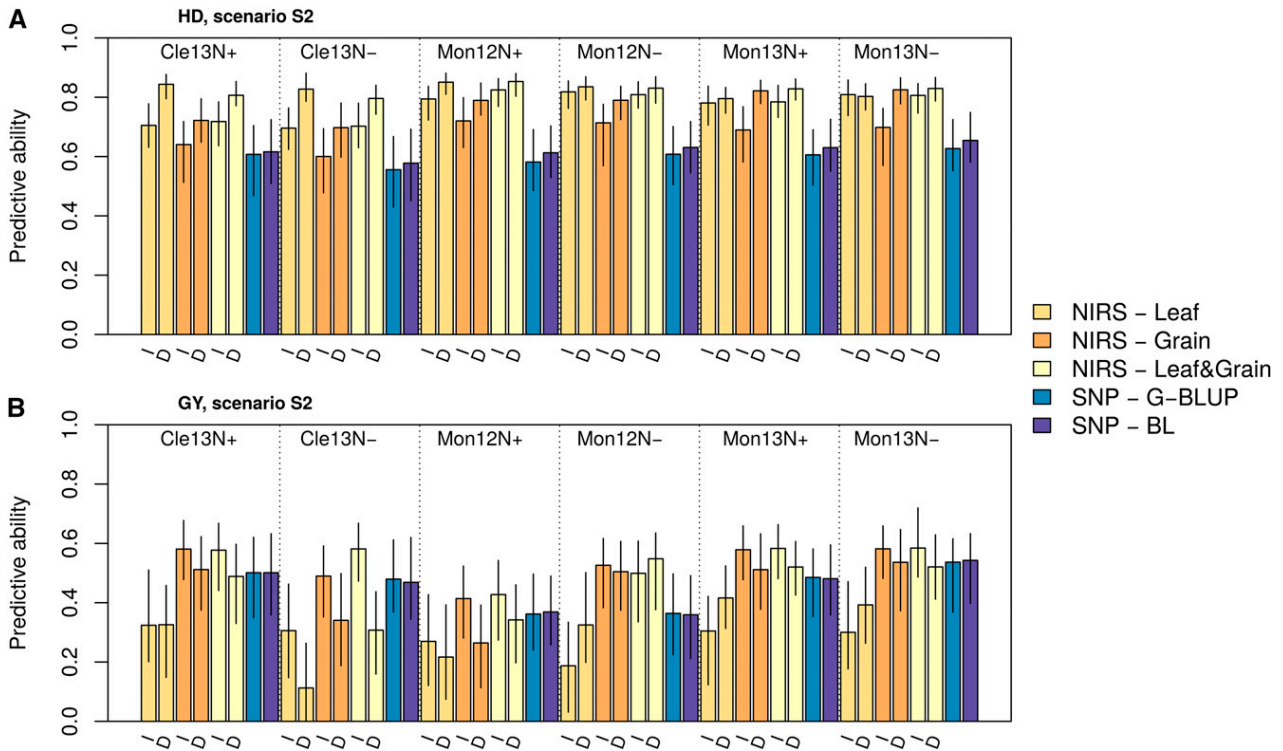


Figure 4 Details of the predictive abilities obtained in scenario S2 for heading date (a) and grain yield (b) for wheat. In S2, the NIRS and phenotypic data used to train the RR-BLUP model were collected in distinct environments. The bars are labeled with the origin of the NIRS data (I: irrigated treatment, D: drought treatment). The medians of the predictive abilities obtained over repeated cross-validations are reported as the height of the bars together with the first and third quartiles as confidence intervals.

across multiple generations. Nevertheless on the other side it can be highly valuable to predict phenotypes, including the effect of interactions and regulatory networks, at key steps of the breeding schemes. In plant breeding, one major objective during the first few generations is to produce numerous individuals with the same genotypes (by self-fertilization,

doubled haploid techniques or clonal reproduction) to allow for field evaluation in multi-environment trial (MET). Because this field evaluation is the most expensive step in the breeding schemes and because it is applied on replicable genotypes, PS would be of major interest to select among all candidates the genotypes that will be evaluated in the MET. In this

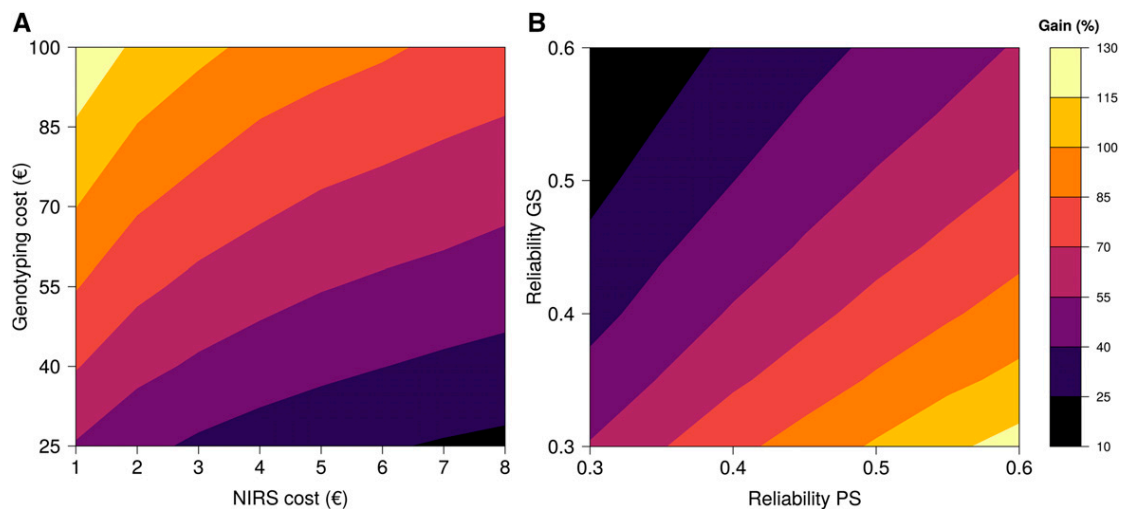


Figure 5 Theoretical increase of expected genetic gain (%) by using NIRS instead of genotyping. a: Expected genetic gain for various genotyping and NIRS costs for a reliability of 0.4, a budget of 200,000 €, and a selection of 400 individuals. b: Expected genetic gain for various reliabilities, a budget of 200,000 €, genotyping and NIRS costs of 50 € and 4 €, respectively, and a selection of 400 individuals. For each scenario, true breeding values and estimated breeding values were simulated thanks to multivariate normal distributions with a covariance adapted to the chosen reliability.

situation, predicting phenotypes instead of additive values is clearly an advantage as the same genotypes can be replicated many times. Another important question related to the efficiency of PS across multiple generations is about the frequency of formula update to maintain a sufficient level of accuracy. This question, which is also relevant for GS, must be addressed in future work. We also believe that future work should assess the efficiency of PS with NIRS obtained from tissues collected on young plants. Ideally, we would like PS to be efficient with NIRS collected on the youngest possible plant to have the information as early as possible (similarly to what can be achieved with GS) and at low cost. We could show for wheat that for fixed material, NIRS collected on seeds, so before sowing, was efficient to run PS, which offers very interesting perspectives for this species. The studies on endophenotypic variations in maize (Fu *et al.* 2012; Riedelsheimer *et al.* 2012; Guo *et al.* 2016; Schrag *et al.* 2018), rice (Xu *et al.* 2016) and wheat (Ward *et al.* 2015) also demonstrated that the characterization of germinated seeds or seedlings was efficient to estimate kinships resulting in accurate predictions. These results are promising, but this needs to be tested for other species and on other datasets. The choice of the tissue on which to collect NIRS is also important in regard to the amount of work required. It is clear that it would be much easier to take NIRS on seeds in the lab than on individual leaves in the field. Further work is necessary to evaluate the feasibility of PS in breeding schemes. One last important point that needs to be addressed in future work, is the way NIRS collected in different environments should be combined to run PS. In the scenarios that we described here, the NIRS of the calibration set and the predicted set were obtained in a same environment. But in practice, breeders collect NIRS each year on different materials, and so it would be necessary to combine NIRS obtained in different environments, which could reduce the prediction accuracy of PS. There are practical ways (such as repeating checks each year, or implementing management practices to homogenize the environments) and theoretical ways (such as Single Step GBLUP (Legarra *et al.* 2009)) to deal with this issue, but these need to be tested.

There are various applications of PS, which we see both as a complement and as an alternative to GS depending on the situation. The first obvious application of PS is its use when no genotyping tool is available at a reasonable cost, which is still the case for many orphan organisms, even if important progress has been achieved thanks to genotyping by sequencing that makes genotyping more accessible and cost effective (Elshire *et al.* 2011; Gorjanc *et al.* 2017a). For these species, PS could potentially be a new efficient breeding tool to increase genetic progress. As mentioned before, a second application would be to use PS to screen nearly fixed material or clones, as PS (in the same manner as selection on endophenotypes) is likely to capture non-additive genetic effects. Even if the prediction accuracy is low, PS can be used to filter out a given proportion of selection candidates. One should define this proportion with respect to PS accuracy: the higher the accuracy, the more confident we are at filtering out many individuals without losing the best candidates. Note that even if PS is less accurate than GS, it could nevertheless be interesting to filter out the worst individuals considering the low cost of NIRS acquisition, and the fact that NIRS is often already routinely carried out (for example, in cereals or forest trees to predict quality traits). In a second step, one could use GS to make complementary predictions on a limited number of selection candidates. A last major application of PS would be to help conservation geneticists manage diversity collections. The use of genotyping to organize seed banks and to screen and define core collections is strongly limited by its cost. PS offers a new opportunity to manage seed banks because it allows distance matrices to be computed cheaply and reliably.

Considering that PS gave interesting results for both a tree and an annual crop regarding various traits related to development,

productivity and tolerance to disease and using tissues of a completely different nature (wood, leaf, grain), we can expect PS to work in many other plants and possibly in animal species using NIRS on organic tissues or fluids. Our work constitutes a proof of concept and a first attempt at PS, which clearly opens new perspectives for the breeding community. Indeed, one could further optimize many parameters to increase PS efficiency. The differences observed here between the PS efficiencies reported for wheat and poplar could represent a first direction for improving the approach. Indeed, PS appeared to be more efficient in wheat than in poplar and several hypotheses could be proposed to explain this result. First, spectra were acquired on different spectrometers resulting in a broader wavelength range in wheat, which also covered the visible part of the electromagnetic spectrum. Consequently, the information brought by the spectra on wheat tissues was potentially richer than the one brought by the spectra on poplar. Second, we could see that in wheat a larger proportion of G and $G \times E$ variance could be captured by the spectra regardless of the tissue sampled and that this was especially true for the lowest wavelengths (including the visible part), which were absent in poplar. Third, the tissues in which NIRS was collected differed, and this difference seems to be an important parameter as highlighted by the differences in predictive ability between leaf and grain in wheat.

Another possibility for the improvement of PS efficiency could be the optimization of the growing conditions of plants in the reference experiment. In wheat, it was typically better to use NIRS collected on plants grown in unfavorable conditions than in favorable conditions. This result might be explained by more pronounced dissimilarities between genetically distant individuals in conditions of stress. Therefore, there is a clear need to optimize these conditions. Once the NIRS data are collected, one could also try to improve the pretreatment of the signal and the statistical model of calibration. In our case, we choose as pretreatment the first derivative of the normalized spectrum, but other options could be tested, and these options might not necessarily be the same depending on the species considered, environment, tissue sampled or target trait. For calibrations, we have used RR-BLUP, but one might test other techniques, such as those typically allowing non-additive effects or involving feature selection, to improve the accuracy of PS. These points clearly indicate that there is great room of improvement of PS, which will likely constitute in the near future an active field of research. Finally, the recent advent of portable NIR devices as well as of hyperspectral imaging allows this technology to be used in the field. Unmanned vehicles and robots are currently being developed and can already be used to automatically collect reflectance at an industrial scale (Madec *et al.* 2017; Aguete *et al.* 2017). These new developments will considerably increase the throughput and conversely decrease the cost of NIRS data. We thus expect that these technological advances will reinforce the advantages of the proposed PS. However, whether these technological advances will have the same predictive ability as NIRS remains to be tested and is likely to be the subject of active research in the near future.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the staff of the INRA GBFOR experimental unit for the establishment and management of the poplar experimental design in Orléans, the collection of wood samples in each site, and their contribution to phenotypic measurements on poplars in Orléans; Alasia Franco Vivai staff for management of the poplar experimental plantation in Savigliano, and M. Sabatti and F. Fabbri for their contribution to phenotypic measurements on poplars in Savigliano. We acknowledge the staff of the INRA Geno-Bois platform for the preparation of samples and collection of NIRS

on wood samples and the staff of EPGV and BioForA for their contribution to obtaining SNP data on poplar. We would like to thank J. Messaoud for NIRS acquisition on wheat samples, V. Allard, B. Adam and D. Cormier for implementation of the rain-out shelter experiment (Phéno3C, INRA Clermont-Ferrand), and E. Heumez (UE GCIE) for the experiment in Estrées-Mons. We would also like to thank A. Chateigner, L. Sanchez, G. Charmet, V. Allard, P. Martre, L. Inchboard and S. Bouchet for useful discussions and comments on the manuscript. We are grateful to the Genotoul bioinformatics platform Toulouse Midi-Pyrenees (Bioinfo Genotoul) for providing computing resources. Establishment and management of the poplar experimental sites until harvests were carried out with financial support from the NOVELTREE project (EU-FP7-211868). NIRS measurements on poplar wood samples were supported by the SYBIOPOP project funded by the French National Research Agency (ANR-13-JSV6-0001). Management of the wheat multi-environment trials was financially supported by the French National Research National Agency under Investment for the Future (BreedWheat project ANR-10-BTBR-03) and by FranceAgriMer. The Phéno3C platform was financially funded by the French National Research National Agency under the Investment for the Future Phenome project (ANR-11-INBS-12) and by the European Regional Development Fund (AV0011535).

V.S. and R.R. designed the study, analyzed the data and wrote the paper with input from J-P.C., P.F.R., E.P., J.L.G., and C.B.

LITERATURE CITED

- Abasolo, M., D. J. Lee, C. Raymond, R. Meder, and M. Shepherd, 2013 Deviant near-infrared spectra identifies *Corymbia* hybrids. *For. Ecol. Manage.* 304: 121–131. <https://doi.org/10.1016/j.foreco.2013.04.040>
- Adedipe, O. E., B. Dawson-Andoh, J. Slahor, and L. Osborn, 2008 Classification of red oak (*Quercus rubra*) and white oak (*Quercus alba*) wood using a near infrared spectrometer and soft independent modelling of class analogies. *J. Near Infrared Spectrosc.* 16: 49–57. <https://doi.org/10.1255/jnirs.760>
- Aguate, F. M., S. Trachsel, L. G. Pérez, J. Burgueño, J. Crossa *et al.*, 2017 Use of Hyperspectral Image Data Outperforms Vegetation Indices in Prediction of Maize Yield. *Crop Sci.* 57: 2517. <https://doi.org/10.2135/cropsci2017.01.0007>
- Bertrand, D., P. Robert, and W. Loisel, 1985 Identification of some wheat varieties by near infrared reflectance spectroscopy. *J. Sci. Food Agric.* 36: 1120–1124. <https://doi.org/10.1002/jsfa.2740361114>
- Cabrera-Bosquet, L., J. Crossa, J. von Zitzewitz, M. D. Serret, and J. Luis Araus, 2012 High-throughput Phenotyping and Genomic Selection: The Frontiers of Crop Breeding Converge. *J. Integr. Plant Biol.* 54: 312–320. <https://doi.org/10.1111/j.1744-7909.2012.01116.x>
- Covarrubias-Pazarán, G., 2016 Genome-Assisted Prediction of Quantitative Traits Using the R Package sommer. *PLoS One* 11: e0156744. <https://doi.org/10.1371/journal.pone.0156744>
- Crossa, J., D. Jarquín, J. Franco, P. Pérez-Rodríguez, J. Burgueño *et al.*, 2016 Genomic Prediction of Gene Bank Wheat Landraces. *G3: Genes|Genomes|Genetics* (Bethesda) 6: 1819–1834. <https://doi.org/10.1534/g3.116.029637>
- Cruikshank, R. H., and L. Munck, 2011 It's barcoding Jim, but not as we know it. *Zootaxa* 56: 55–56.
- de los Campos, G., J. M. Hickey, R. Pong-Wong, H. D. Daetwyler, and M. P. L. Calus, 2013 Whole-Genome Regression and Prediction Methods Applied to Plant and Animal Breeding. *Genetics* 193: 327–345. <https://doi.org/10.1534/genetics.112.143313>
- de los Campos, G., H. Naya, D. Gianola, J. Crossa, A. Legarra *et al.*, 2009 Predicting Quantitative Traits With Regression Models for Dense Molecular Markers and Pedigree. *Genetics* 182: 375–385. <https://doi.org/10.1534/genetics.109.101501>
- Diago, M. P., A. M. Fernandes, B. Millan, J. Tardaguila, and P. Melo-Pinto, 2013 Identification of grapevine varieties using leaf spectroscopy and partial least squares. *Comput. Electron. Agric.* 99: 7–13. <https://doi.org/10.1016/j.compag.2013.08.021>
- Diepeveen, D., G. P. Y. Clarke, K. Ryan, A. Tarr, W. Ma *et al.*, 2012 Molecular genetic mapping of NIR spectra variation. *J. Cereal Sci.* 55: 6–14.
- Dillen, S. Y., N. Marron, M. Sabatti, R. Ceulemans, and C. Bastien, 2009 Relationships among productivity determinants in two hybrid poplar families grown during three years at two contrasting sites. *Tree Physiol.* 29: 975–987. <https://doi.org/10.1093/treephys/tpp036>
- Ecarnot, M., F. Compan, and P. Roumet, 2013 Assessing leaf nitrogen content and leaf mass per unit area of wheat in the field throughout plant cycle with a portable spectrometer. *Field Crops Res.* 140: 44–50. <https://doi.org/10.1016/j.fcr.2012.10.013>
- Elshire, R. J., J. C. Glaubitz, Q. Sun, J. A. Poland, K. Kawamoto *et al.*, 2011 A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One* 6: e19379. <https://doi.org/10.1371/journal.pone.0019379>
- Endelman, J. B., 2011 Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. *The Plant Genome Journal* 4: 250. <https://doi.org/10.3835/plantgenome2011.08.0024>
- Espinoza, J. A., G. R. Hodge, and W. S. Dvorak, 2012 The potential use of near infrared spectroscopy to discriminate between different pine species and their hybrids. *J. Near Infrared Spectrosc.* 20: 437–447. <https://doi.org/10.1255/jnirs.1006>
- Faivre-Rampant, P., G. Zaina, V. Jorge, S. Giacomello, V. Segura *et al.*, 2016 New resources for genetic studies in *Populus nigra*: genome-wide SNP discovery and development of a 12k Infinium array. *Mol. Ecol. Resour.* 16: 1023–1036. <https://doi.org/10.1111/1755-0998.12513>
- Feher, K., J. Liseč, L. Römisch-Margl, J. Selbig, A. Gierl *et al.*, 2014 Deducing Hybrid Performance from Parental Metabolic Profiles of Young Primary Roots of Maize by Using a Multivariate Diallel Approach. *PLoS One* 9: e85435. <https://doi.org/10.1371/journal.pone.0085435>
- Fernandez, O., M. Urrutia, S. Bernillon, C. Giauffret, F. Tardieu *et al.*, 2016 Fortune telling: metabolic markers of plant performance. *Metabolomics* 12: 158. <https://doi.org/10.1007/s11306-016-1099-1>
- Ferrio, J., D. Villegas, J. Zarco, N. Aparicio, J. Araus *et al.*, 2005 Assessment of durum wheat yield using visible and near-infrared reflectance spectra of canopies. *Field Crops Res.* 94: 126–148. <https://doi.org/10.1016/j.fcr.2004.12.002>
- Fischnaller, S., F. E. Dowell, A. Lusser, B. C. Schlick-Steiner, and F. M. Steiner, 2012 Non-destructive species identification of *Drosophila obscura* and *D. subobscura* (Diptera) using near-infrared spectroscopy. *Fly (Austin)* 6: 284–289. <https://doi.org/10.4161/fly.21535>
- Foley, W. J., A. McIlwee, I. Lawler, L. Aragones, A. P. Woolnough *et al.*, 1998 Ecological applications of near infrared reflectance spectroscopy - a tool for rapid, cost-effective prediction of the composition of plant and animal tissues and aspects of animal performance. *Oecologia* 116: 293–305. <https://doi.org/10.1007/s004420050591>
- Forni, S., I. Aguilar, and I. Misztal, 2011 Different genomic relationship matrices for single-step analysis using phenotypic, pedigree and genomic information. *Genet. Sel. Evol.* 43: 1. <https://doi.org/10.1186/1297-9686-43-1>
- Fu, J., K. C. Falke, A. Thiemann, T. A. Schrag, A. E. Melchinger *et al.*, 2012 Partial least squares regression, support vector machine regression, and transcriptome-based distances for prediction of maize hybrid performance with gene expression data. *Theor. Appl. Genet.* 124: 825–833. <https://doi.org/10.1007/s00122-011-1747-9>
- Gaynor, R. C., G. Gorjanc, A. R. Bentley, E. S. Ober, P. Howell *et al.*, 2017 A Two-Part Strategy for Using Genomic Selection to Develop Inbred Lines. *Crop Sci.* 57: 2372. <https://doi.org/10.2135/cropsci2016.09.0742>
- Gebreselassie, M. N., K. Ader, N. Boizot, F. Millier, J.-P. P. Charpentier *et al.*, 2017 Near-infrared spectroscopy enables the genetic analysis of chemical properties in a large set of wood samples from *Populus nigra* (L.) natural populations. *Ind. Crops Prod.* 107: 159–171.
- Genz, A., F. Bretz, T. Miwa, X. Mi, F. Leisch, *et al.*, 2018 *mvtnorm: Multivariate Normal and t Distributions*. R package version 1.0–8.

- Goddard, M., 2009 Genomic selection: Prediction of accuracy and maximisation of long term response. *Genetica* 136: 245–257. <https://doi.org/10.1007/s10709-008-9308-0>
- Gorjanc, G., M. Battagin, J.-F. Dumasy, R. Antolin, R. C. Gaynor *et al.*, 2017a Prospects for Cost-Effective Genomic Selection via Accurate Within-Family Imputation. *Crop Sci.* 57: 216. <https://doi.org/10.2135/cropsci2016.06.0526>
- Gorjanc, G., J.-F. Dumasy, S. Gonen, R. C. Gaynor, R. Antolin *et al.*, 2017b Potential of Low-Coverage Genotyping-by-Sequencing and Imputation for Cost-Effective Genomic Selection in Biparental Segregating Populations. *Crop Sci.* 57: 1404. <https://doi.org/10.2135/cropsci2016.08.0675>
- Gorjanc, G., J. Jenko, S. J. Hearne, and J. M. Hickey, 2016 Initiating maize pre-breeding programs using genomic selection to harness polygenic variation from landrace populations. *BMC Genomics* 17: 30. <https://doi.org/10.1186/s12864-015-2345-z>
- Guet, J., F. Fabbrini, R. Fichot, M. Sabatti, C. Bastien *et al.*, 2015 Genetic variation for leaf morphology, leaf structure and leaf carbon isotope discrimination in European populations of black poplar (*Populus nigra* L.). *Tree Physiol.* 35: 850–863. <https://doi.org/10.1093/treephys/tpv056>
- Guo, Z., M. M. Magwire, C. J. Basten, Z. Xu, and D. Wang, 2016 Evaluation of the utility of gene expression and metabolic information for genomic prediction in maize. *Theor. Appl. Genet.* 129: 2413–2427. <https://doi.org/10.1007/s00122-016-2780-5>
- Habier, D., R. L. Fernando, and J. C. M. Dekkers, 2007 The impact of genetic relationship information on genome-assisted breeding values. *Genetics* 177: 2389–2397. <https://doi.org/10.1534/genetics.107.081190>
- Hayes, B., P. M. Visscher, and M. Goddard, 2009 Increased accuracy of artificial selection by using the realized relationship matrix. *Genet. Res.* 91: 47. <https://doi.org/10.1017/S0016672308009981>
- Heffner, E. L., A. J. Lorenz, J.-L. Jannink, and M. E. Sorrells, 2010 Plant Breeding with Genomic Selection: Gain per Unit Time and Cost. *Crop Sci.* 50: 1681. <https://doi.org/10.2135/cropsci2009.11.0662>
- Hein, P. R. G., and G. Chaix, 2014 NIR Spectral Heritability: A Promising Tool for Wood Breeders? *J. Near Infrared Spectrosc.* 22: 141–147. <https://doi.org/10.1255/jnirs.1108>
- Kang, H. M., J. H. Sul, S. K. Service, N. A. Zaitlen, S.-Y. Kong *et al.*, 2010 Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* 42: 348–354. <https://doi.org/10.1038/ng.548>
- Korte, A., B. J. Vilhjálmsson, V. Segura, A. Platt, Q. Long *et al.*, 2012 A mixed-model approach for genome-wide association studies of correlated traits in structured populations. *Nat. Genet.* 44: 1066–1071. <https://doi.org/10.1038/ng.2376>
- Lang, C., D. R. Almeida, and F. R. Costa, 2017 Discrimination of taxonomic identity at species, genus and family levels using Fourier Transformed Near-Infrared Spectroscopy (FT-NIR). *For. Ecol. Manage.* 406: 219–227. <https://doi.org/10.1016/j.foreco.2017.09.003>
- Legarra, A., I. Aguilar, and I. Misztal, 2009 A relationship matrix including full pedigree and genomic information. *J. Dairy Sci.* 92: 4656–4663. <https://doi.org/10.3168/jds.2009-2061>
- Ly, D., S. Huet, A. Gauffreteau, R. Rincint, G. Touzy *et al.*, 2018 Whole-genome prediction of reaction norms to environmental stress in bread wheat (*Triticum aestivum* L.) by genomic random regression. *Field Crops Res.* 216: 32–41. <https://doi.org/10.1016/j.fcr.2017.08.020>
- Mackay, T. F. C., E. a. Stone, and J. F. Ayroles, 2009 The genetics of quantitative traits: challenges and prospects. *Nat. Rev. Genet.* 10: 565–577. <https://doi.org/10.1038/nrg2612>
- Madedec, S., F. Baret, B. de Solan, S. Thomas, D. Dutartre *et al.*, 2017 High-Throughput Phenotyping of Plant Height: Comparing Unmanned Aerial Vehicles and Ground LiDAR Estimates. *Front. Plant Sci.* 8: 2002. <https://doi.org/10.3389/fpls.2017.02002>
- Meder, R., D. Kain, N. Ebdon, P. Macdonell, and J. T. Brawner, 2014 Identifying hybridisation in *Pinus* species using near infrared spectroscopy of foliage. *J. Near Infrared Spectrosc.* 22: 337–345.
- Meuwissen, T., B. Hayes, and M. Goddard, 2013 Accelerating Improvement of Livestock with Genomic Selection. *Annu. Rev. Anim. Biosci.* 1: 221–237. <https://doi.org/10.1146/annurev-animal-031412-103705>
- Meuwissen, T. H. E., B. J. Hayes, and M. E. Goddard, 2001 Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157: 1819–1829.
- Muñoz, F. and L. Sanchez, 2017 *breedR: Statistical Methods for Forest Genetic Resources Analysts*. R package version 0.12–2.
- Oakey, H., A. P. Verbyla, B. R. Cullis, X. Wei, and W. S. Pitchford, 2007 Joint modeling of additive and non-additive (genetic line) effects in multi-environment trials. *Theor. Appl. Genet.* 114: 1319–1332. <https://doi.org/10.1007/s00122-007-0515-3>
- O'Reilly-Wapstra, J. M., J. S. Freeman, R. Barbour, R. E. Vaillancourt, and B. M. Potts, 2013 Genetic analysis of the near-infrared spectral phenotype of a global *Eucalyptus* species. *Tree Genet. Genomes* 9: 943–959. <https://doi.org/10.1007/s11295-013-0607-y>
- Park, T., and G. Casella, 2008 The Bayesian Lasso. *J. Am. Stat. Assoc.* 103: 681–686. <https://doi.org/10.1198/01621450800000337>
- Peerbhay, K. Y., O. Mutanga, and R. Ismail, 2013 Commercial tree species discrimination using airborne AISA Eagle hyperspectral imagery and partial least squares discriminant analysis (PLS-DA) in KwaZulu-Natal, South Africa. *ISPRS J. Photogramm. Remote Sens.* 79: 19–28. <https://doi.org/10.1016/j.isprsjprs.2013.01.013>
- Posada, H., M. Ferrand, F. Davrieux, P. Lashermes, and B. Bertrand, 2009 Stability across environments of the coffee variety near infrared spectral signature. *Heredity* 102: 113–119. <https://doi.org/10.1038/hdy.2008.88>
- R Core Team, 2018 *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
- Riedelshheimer, C., A. Czedik-Eysenberg, C. Grieder, J. Lisec, F. Technow *et al.*, 2012 Genomic and metabolic prediction of complex heterotic traits in hybrid maize. *Nat. Genet.* 44: 217–220. <https://doi.org/10.1038/ng.1033>
- Rimbert, H., B. Darrier, J. Navarro, J. Kitt, F. Choulet *et al.*, 2018 High throughput SNP discovery and genotyping in hexaploid wheat. *PLoS One* 13: e0186329. <https://doi.org/10.1371/journal.pone.0186329>
- Rodríguez-Álvarez, M. X., M. P. Boer, F. A. van Eeuwijk, and P. H. Eilers, 2017 Correcting for spatial heterogeneity in plant breeding experiments with p-splines. *Spat. Stat.* 23: 52–71. <https://doi.org/10.1016/j.spasta.2017.10.003>
- Rohde, A., V. Storme, V. Jorge, M. Gaudet, N. Vitacolonna *et al.*, 2011 Bud set in poplar - genetic dissection of a complex trait in natural and hybrid populations. *New Phytol.* 189: 106–121. <https://doi.org/10.1111/j.1469-8137.2010.03469.x>
- Sargolzaei, M., J. P. Chesnais, and F. S. Schenkel, 2014 A new approach for efficient genotype imputation using information from relatives. *BMC Genomics* 15: 478. <https://doi.org/10.1186/1471-2164-15-478>
- Savitzky, A., and M. J. E. Golay, 1964 Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Anal. Chem.* 36: 1627–1639. <https://doi.org/10.1021/ac60319a045>
- Schrag, T. A., M. Westhues, W. Schipprack, F. Seifert, A. Thiemann *et al.*, 2018 Beyond Genomic Prediction: Combining Different Types of omics-Data Can Improve Prediction of Hybrid Performance in Maize. *Genetics* 208: genetics.300374.2017. <https://doi.org/10.1534/genetics.117.300374>
- Seifert, F., A. Thiemann, T. A. Schrag, D. Rybka, A. E. Melchinger *et al.*, 2018 Small RNA-based prediction of hybrid performance in maize. *BMC Genomics* 19: 371. <https://doi.org/10.1186/s12864-018-4708-8>
- signal developers, 2014 *signal: Signal processing*.
- Teixeira Dos Santos, C. A., M. Lopo, R. N. M. J. Páscoa, and J. A. Lopes, 2013 A review on the applications of portable near-infrared spectrometers in the agro-food industry. *Appl. Spectrosc.* 67: 1215–1233. <https://doi.org/10.1366/13-07228>
- VanRaden, P. M., 2008 Efficient methods to compute genomic predictions. *J. Dairy Sci.* 91: 4414–4423. <https://doi.org/10.3168/jds.2007-0980>
- Ward, J., M. Rakszegi, Z. Bedő, P. R. Shewry, and I. Mackay, 2015 Differentially penalized regression to predict agronomic traits from metabolites and markers in wheat. *BMC Genet.* 16: 19. <https://doi.org/10.1186/s12863-015-0169-0>
- Weber, V., J. Araus, J. Cairns, C. Sanchez, A. Melchinger *et al.*, 2012 Prediction of grain yield using reflectance spectra of canopy and

- leaves in maize plants grown under different water regimes. *Field Crops Res.* 128: 82–90. <https://doi.org/10.1016/j.fcr.2011.12.016>
- Westhues, M., T. A. Schrag, C. Heuer, G. Thaller, H. F. Utz *et al.*, 2017 Omics-based hybrid prediction in maize. *Theor. Appl. Genet.* 130: 1927–1939. <https://doi.org/10.1007/s00122-017-2934-0>
- Whittaker, J., 2000 Marker-assisted selection using ridge regression. *Genetical* 75: 351–367.
- Xu, S., Y. Xu, L. Gong, and Q. Zhang, 2016 Metabolomic prediction of yield in hybrid rice. *Plant J.* 88: 219–227. <https://doi.org/10.1111/tpj.13242>
- Yamada, Y., Y. Itoh, and I. Sugimoto, 1988 Parametric relationships between genotype x environment interaction and genetic correlation when two environments are involved. *Theor. Appl. Genet.* 76: 850–854. <https://doi.org/10.1007/BF00273671>
- Yu, X., X. Li, T. Guo, C. Zhu, Y. Wu *et al.*, 2016 Genomic prediction contributing to a promising global strategy to turbocharge gene banks. *Nat. Plants* 2: 16150. <https://doi.org/10.1038/nplants.2016.150>
- Zenke-Philippi, C., M. Frisch, A. Thiemann, F. Seifert, T. Schrag *et al.*, 2017 Transcriptome-based prediction of hybrid performance with unbalanced data from a maize breeding programme. *Plant Breed.* 136: 331–337. <https://doi.org/10.1111/pbr.12482>
- Zhong, S., J. C. M. Dekkers, R. L. Fernando, and J.-L. Jannink, 2009 Factors Affecting Accuracy From Genomic Selection in Populations Derived From Multiple Inbred Lines: A Barley Case Study. *Genetics* 182: 355–364. <https://doi.org/10.1534/genetics.108.098277>


Communicating editor: J. Holland

RESEARCH ARTICLE

Open Access

Gene expression predictions and networks in natural populations supports the omnigenic theory



Aurélien Chateigner¹, Marie-Claude Lesage-Descauses¹, Odile Rogier¹, Véronique Jorge¹, Jean-Charles Leplé², Véronique Brunaud^{3,4}, Christine Paysant-Le Roux^{3,4}, Ludivine Soubigou-Taconnat^{3,4}, Marie-Laure Martin-Magniette^{3,4,5}, Leopoldo Sanchez^{1†} and Vincent Segura^{1,6*} 

Abstract

Background: Recent literature on the differential role of genes within networks distinguishes core from peripheral genes. If previous works have shown contrasting features between them, whether such categorization matters for phenotype prediction remains to be studied.

Results: We measured 17 phenotypic traits for 241 cloned genotypes from a *Populus nigra* collection, covering growth, phenology, chemical and physical properties. We also sequenced RNA for each genotype and built co-expression networks to define core and peripheral genes. We found that cores were more differentiated between populations than peripherals while being less variable, suggesting that they have been constrained through potentially divergent selection. We also showed that while cores were overrepresented in a subset of genes statistically selected for their capacity to predict the phenotypes (by Boruta algorithm), they did not systematically predict better than peripherals or even random genes.

Conclusion: Our work is the first attempt to assess the importance of co-expression network connectivity in phenotype prediction. While highly connected core genes appear to be important, they do not bear enough information to systematically predict better quantitative traits than other gene sets.

Keywords: Core, Peripheral, Boruta, Machine learning, *Populus nigra*

Background

Gene-to-gene interaction is a pervasive although elusive phenomenon underlying phenotype expression. Genes operate within networks with more or less mediated actions on the phenome. Systems biology approaches are required to grasp the functional topology of these networks and ultimately gain insights into how gene interactions interplay at different biological levels to produce

global phenotypes [1]. New sources of information and their subsequent use in the inference of gene networks are populating the wide gap existing between phenotypes and DNA sequences and, therefore, opening the door to systems biology approaches for the development of context-dependent phenotypic predictions. RNA sequencing (RNA-seq) is one of such new sources of information that can be used to infer gene networks [2].

Among the many works on gene network inference based on transcriptomic data, two recent studies aimed at characterizing the different gene roles within co-expression networks [3, 4]. Josephs et al. [3] studied the link between gene expression, gene connectivity [5], diver-

*Correspondence: vincent.segura@inrae.fr

†Leopoldo Sanchez and Vincent Segura contributed equally to this work.

¹BioForA, INRAE, ONF, Orléans, France

⁶AGAP, Université Montpellier, CIRAD, INRAE, Montpellier SupAgro, Montpellier, France

Full list of author information is available at the end of the article



gence [6] and traces of natural selection [7, 8] in a natural population of the plant *Capsella grandiflora*. They showed that both connectivity and local regulatory variation on the genome are important factors, while not being able to disentangle which of them is directly responsible for patterns of selection among genes. Mähler et al. [4] recalled the importance of studying the general features of biological networks in natural populations. With a genome-wide association study (GWAS) on expression data from RNA-seq, they suggested that purifying selection is the main mechanism maintaining functional connectivity of core genes in a network and that this connectivity is inversely related to eQTLs effect size. These two studies start to outline the first elements of a gene network theory based on connectivity, stating that core genes, which are highly connected, are each of high importance, and thus highly constrained by selection. In contrast to these central genes, there are peripheral, less connected genes, never far from a core hub. These peripheral genes are less constrained than core genes and consequently, they harbor larger amounts of variation at population levels.

Furthermore, classic studies of molecular evolution in biological pathways can help us understand the link between gene connectivity and traits. Several articles showed that selection pressure is correlated to the gene position within the pathway, either positively [9–14] or negatively [9, 15–17], depending on the pathway. Jovelin et al. [15] showed that selective constraints are positively correlated to expression level, confirming previous studies [18–20]. Montanucci et al. [21] showed a positive correlation between selective constraints and connectivity, although such a possibility remained contentious in previous works [22, 23].

While Josephs' [3] and Mähler's [4] studies framed a general view of genes organization based on topological features described in molecular evolution studies of biological pathways, a point remains quite unclear so far: to what extent core and peripheral genes based on connectivity within a co-expression network are involved in the definition of a phenotype? One way to clarify this would be to study the respective roles of core and peripheral genes, as defined on the basis of their connectivity within a co-expression network, in the prediction of a phenotype. Even if predictions are still one step before validation by in vivo experiments, they already represent a landmark that may not only be correlative but also closer to causation, depending on the modeling strategy.

Present study aims at exploring gene ability to predict traits, with datasets representing core genes and peripheral genes, as defined by a topological based model. By making use of two methods to predict phenotypes of available traits, a classic additive linear model, and a more complex and interactive neural network model, we further aimed at studying the mode of action of each type

of genes, in order to gain insight into the genetic architecture of a relatively large range of complex traits. On the one hand, genes that are better predictors with an additive model are supposed to have an overall less redundant, more additive, direct mode of action. On the other hand, genes being better predictors with an interactive model are supposed to operate with high pervasiveness and redundancy, through high connectivity. It is not evident to assign a priori a preferential mode of action and respective roles to core versus peripheral genes. We could assume the former to be downstream genes in biological pathways, closer to the phenotypic expression. The latter could be upstream genes, further away from the phenotype. However, such hypotheses would require levels of data integration that might not be easily available. More readily accessible would be the question of the extent to which connectivity of core genes is captured by models that are sensible to interactivity, involving high but selectively constrained expression levels [15, 21]. With a lower variation, we also expect core genes to be worse predictors for traits than peripheral genes unless the former also bear larger effects.

To answer the questions concerning the respective roles of core and peripheral genes on phenotypic variation, we have sequenced the RNA of 459 samples of black poplar (*Populus nigra*), corresponding to 241 genotypes, from 11 populations representing the natural distribution of the species across Western Europe. We also have, for each of these trees, phenotypic records for 17 traits, covering the growth, phenology, physical and chemical properties of wood. They cover two different environments where the trees were grown in common gardens, in central France and northern Italy. With the transcriptomic data, we built a co-expression network in order to define contrasting gene sets according to their connectivity within the network. We then asked whether these contrasting sets differed in terms of both population and quantitative genetics parameters and quantitative trait prediction.

Results

Wood samples, phenotypes, and transcriptomes

Wood collection and phenotypic data have been previously described [24]. Further details are provided in the “Materials and Methods” section. The complete pipeline is sketched in Fig. 1. Briefly, we are focusing on 241 genotypes coming from different natural populations in western Europe and planted in 2 common gardens (to avoid the confounding between genetic and large environmental effects) at two different locations: Orléans (central France) and Savigliano (northern Italy). Each common garden is composed of 6 replicated and randomized complete blocks. A total of 17 phenotypic traits have been collected on these genotypes (7 traits in common between the two locations, 3 unique to Orléans). These traits could

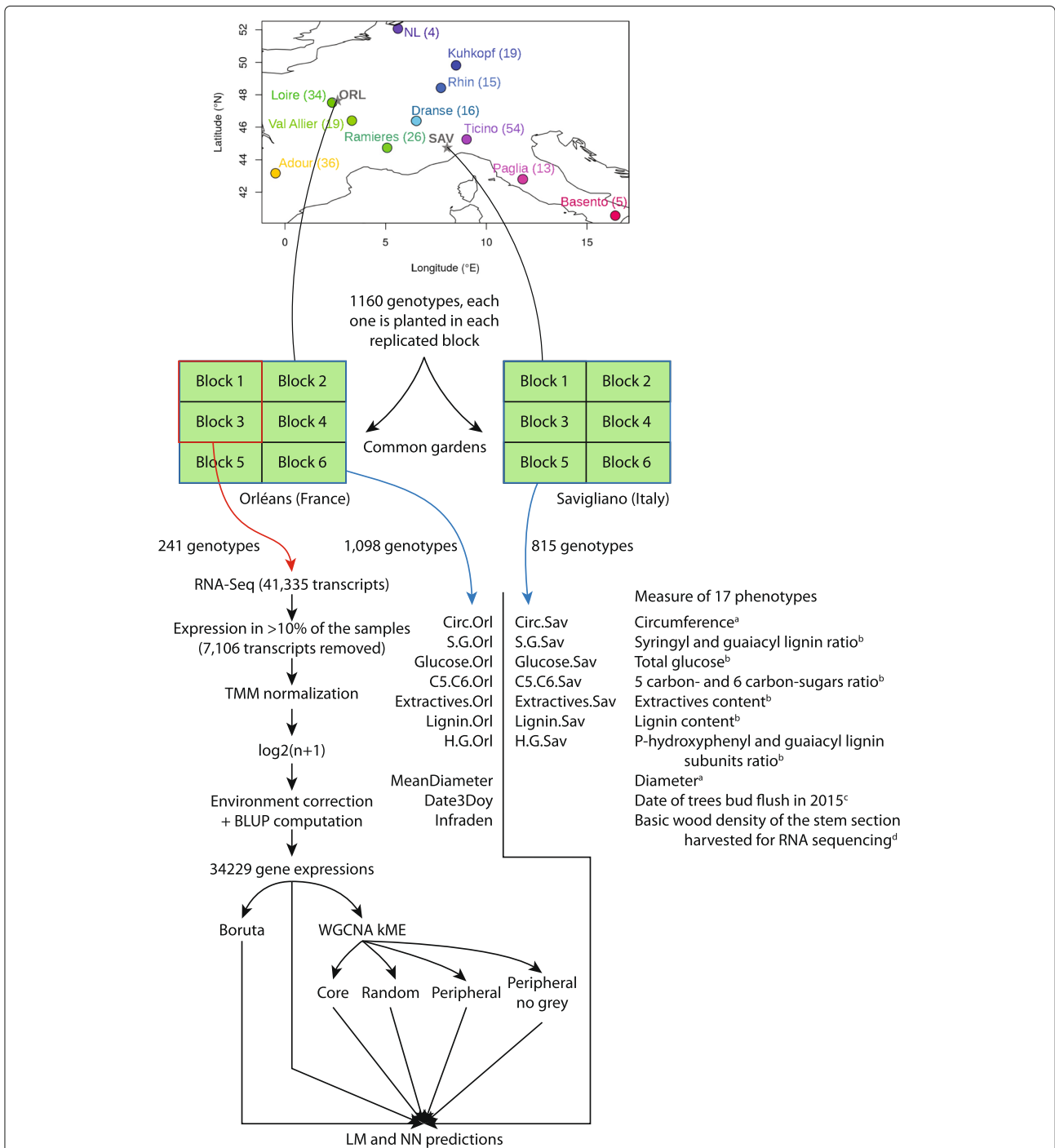


Fig. 1 General sketch of the experiment. From the top to the bottom: Map of the location of the different populations sampled for this experiment, the number of individuals used for the RNA sequencing is indicated between parentheses. From these populations, genotypes were collected and planted in 2 locations (Orléans, in central France, and Savigliano, in northern Italy). At each site, we planted 6 clones of each genotype, 1 in each of the 6 blocks, and their position in each block was randomized. For all the blocks, we collected phenotypes: 10 in Orléans (circumference, S/G, glucose, C5/C6, extractives, lignin, H/G, diameter, infradensity and date of bud flush) and 7 in Savigliano (circumference, S/G, glucose, C5/C6, extractives, lignin, H/G). Only on the clones of 2 blocks in Orléans, we performed the RNA sequencing and treatment of data. The treated RNA-seq data were used with different algorithms and in different sets to predict the phenotypes measured on the same genotype but on different trees (in Savigliano). Trait category: ^aGrowth, ^bChemical, ^cPhenology, ^dPhysical

be organized into four categories depending on the biological process they described (Fig. 1), and they appeared to be quite diverse in terms of genetic control with marker-based heritability estimates ranging between 0.05 and 1 (data not shown). In Orléans only, we used 2 clonal trees per genotype (from 2 blocks) to sample xylem and cambium during the 2015 growing season, and pooled them for RNA sequencing. No tree from Savigliano was used for RNA-seq. Because of sampling and experimental mistakes that were further revealed by the polymorphisms in the RNA sequences, we ended up with 459 samples for which we confirmed the genotype identity (comparison to previously available genotyping data from an SNP chip [25]). These samples corresponded to 218 genotypes with two biological replicates and 23 genotypes with a single biological replicate.

We mapped the sequencing reads on the *Populus trichocarpa* transcriptome (v3.0) to obtain gene expression data. We removed from the data the transcripts for which we did not have at least one count in 10% of the individuals, yielding 34,229 transcripts. We then normalized the data (with TMM) and stabilized the variance (with $\log_2(n + 1)$). RNA collection lasted over a 2-weeks period, with varying weather conditions along the days. We did PCA analyses on the cofactors that were presumably involved in the experience, to look whether any confounding effect could be identified (Suppl. Fig. 1). No clear segregation was found for any of those, except for the ones associated with block, date and hour of sampling. We used a linear mixed-model framework to correct the effects of these cofactors on each transcript (see the “Materials and Methods” section for a formal description of the model used), with R (v3.6.3) [26] and the *breedR* R package (v0.12.2) [27], and further computed from the models the complete BLUP for each genotype. Hereafter, we refer to this set of BLUPs for the 34,229 transcripts as the full gene set (83% of annotated transcripts).

Clustering and network construction

The commonly used approach to build a signed scale-free gene expression network is to use the weighted correlation network analysis (implemented in the WGCNA R package (v1.68) [5]), using a power function on correlations between gene expressions. We chose to use Spearman's rank correlation to avoid any assumption on the linearity of relationships. The scale-free topology fitting index (R^2) did not reach the soft-threshold of 0.85, so we chose the recommended power value of 12, corresponding to the first decrease in the slope growth of the index, resulting in an average connectivity of 195.2 (Fig. 2a). We detected 16 gene expression modules (Suppl. Table 1) with automatic detection (merging threshold: 0.25, minimum module size: 30, Fig. 2b). Spearman correlations between phenotypic and expression data, presented in the lower panel

of Fig. 2b below the module membership of each gene, displayed a structure when the order followed the gene expression tree. The traits themselves were line ordered according to clustering on their scaled values to represent their relationships (Suppl. Fig. 2). Interestingly, most patterns in the correlation between expression and traits did not follow what we would have expected, a certain similarity between sites for a given trait (5 traits with unexpected behavior out of 7 with data in both geographical sites: Circ, S.G., Glucose, Lignin and H.G.). For instance, in the group composed of S/G ratios and glucose composition, the patterns were more similar for different traits in the same site than for the same trait in the different sites (Fig. 2b). Complex shared regulations mediated by the environment seem to be in control of these phenotypes, suggesting site-specific genetic control. Otherwise, glucose composition in Savigliano, wood basic density, and extractives in Orléans presented similar patterns, contrarily to what would be expected from the low phenotypic correlations observed between these traits. These results from the comparative analysis of correlations pinpoint some underlying links between traits that are not obvious from factual phenotypic and genetic correlations between traits.

To get further insight into the relationships between module composition and traits, we looked at the strongest correlations (positive and negative) between the best theoretical representative of a gene expression module (eigengene) and each trait, in order to identify genes in relevant modules with an influence on the trait (Fig. 2c). Following a Bonferroni correction of the p -values provided by WGCNA, only 80 correlations remained significant ($p \leq 0.05$) out of the initial 272 traits by module combinations. Six traits displayed no significant correlations with any module (Glucose.Sav, both C5.C6, Extractives.Sav, Lignin.Sav and H.G.Sav) and 1 module was not significantly correlated with any of the traits studied (purple, Fig. 2c). For those modules showing significant correlations with traits, it was also observed a significant correlation between those expression versus trait correlations and the centrality in the modules (represented by the kME, the correlation with the module eigengene). Conversely, no correlation was found in poorly correlated modules (Fig. 2d, Suppl. Fig. 3). In other words, there was a three-way correlation. The genes with the highest kME in a given module were the most correlated to the eigengene and, consequently, were also the most correlated to those traits with the largest correlation with the module eigengene. Although this is somehow expected, it underlines the usefulness of kME as a centrality score to further characterize the genes within each module. We thus used this centrality score to define further the topological position of our gene expressions in the network and to serve as a basis for role comparisons between

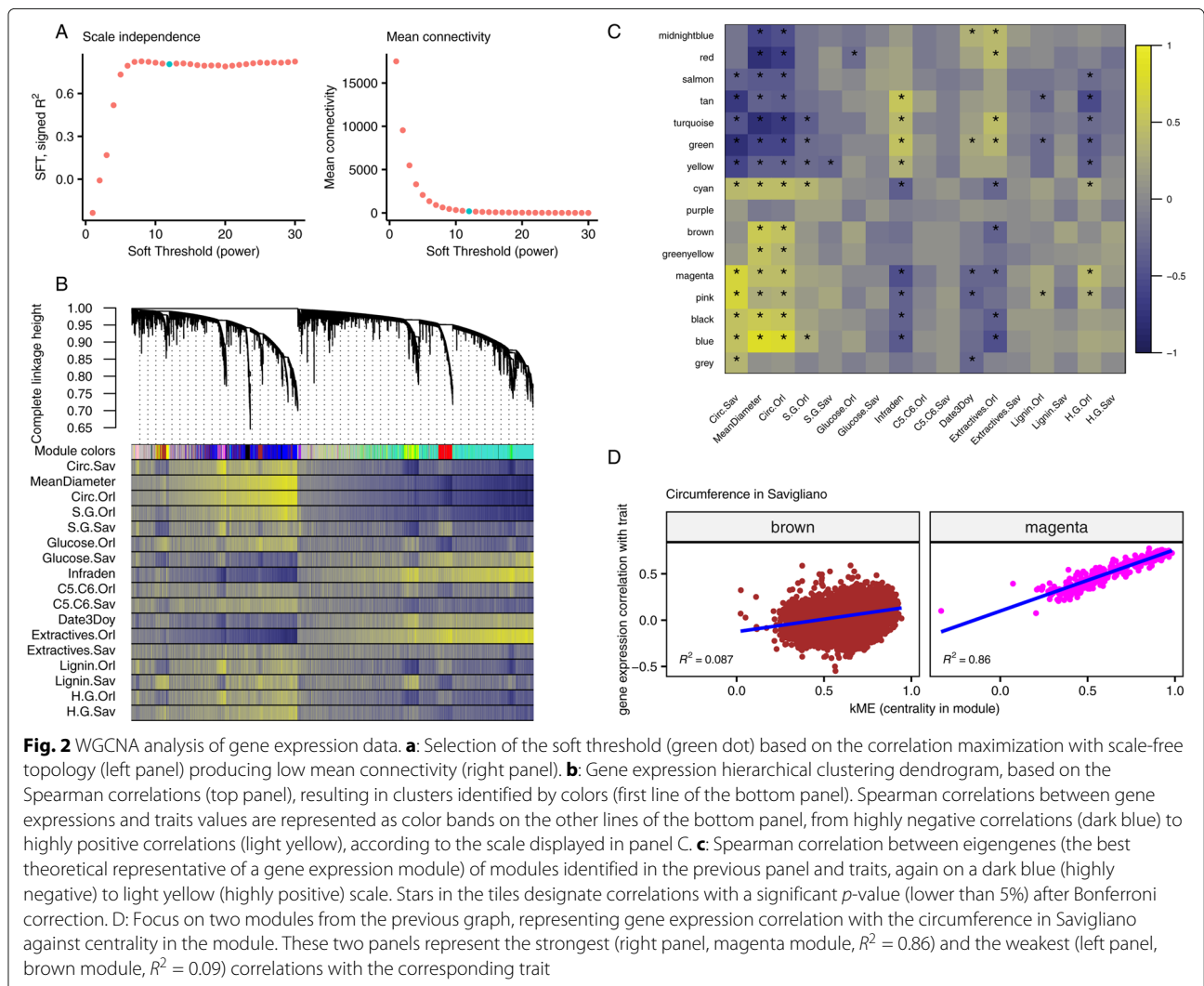


Fig. 2 WGCNA analysis of gene expression data. **a:** Selection of the soft threshold (green dot) based on the correlation maximization with scale-free topology (left panel) producing low mean connectivity (right panel). **b:** Gene expression hierarchical clustering dendrogram, based on the Spearman correlations (top panel), resulting in clusters identified by colors (first line of the bottom panel). Spearman correlations between gene expressions and traits values are represented as color bands on the other lines of the bottom panel, from highly negative correlations (dark blue) to highly positive correlations (light yellow), according to the scale displayed in panel C. **c:** Spearman correlation between eigengenes (the best theoretical representative of a gene expression module) of modules identified in the previous panel and traits, again on a dark blue (highly negative) to light yellow (highly positive) scale. Stars in the tiles designate correlations with a significant p -value (lower than 5%) after Bonferroni correction. **d:** Focus on two modules from the previous graph, representing gene expression correlation with the circumference in Savigliano against centrality in the module. These two panels represent the strongest (right panel, magenta module, $R^2 = 0.86$) and the weakest (left panel, brown module, $R^2 = 0.09$) correlations with the corresponding trait

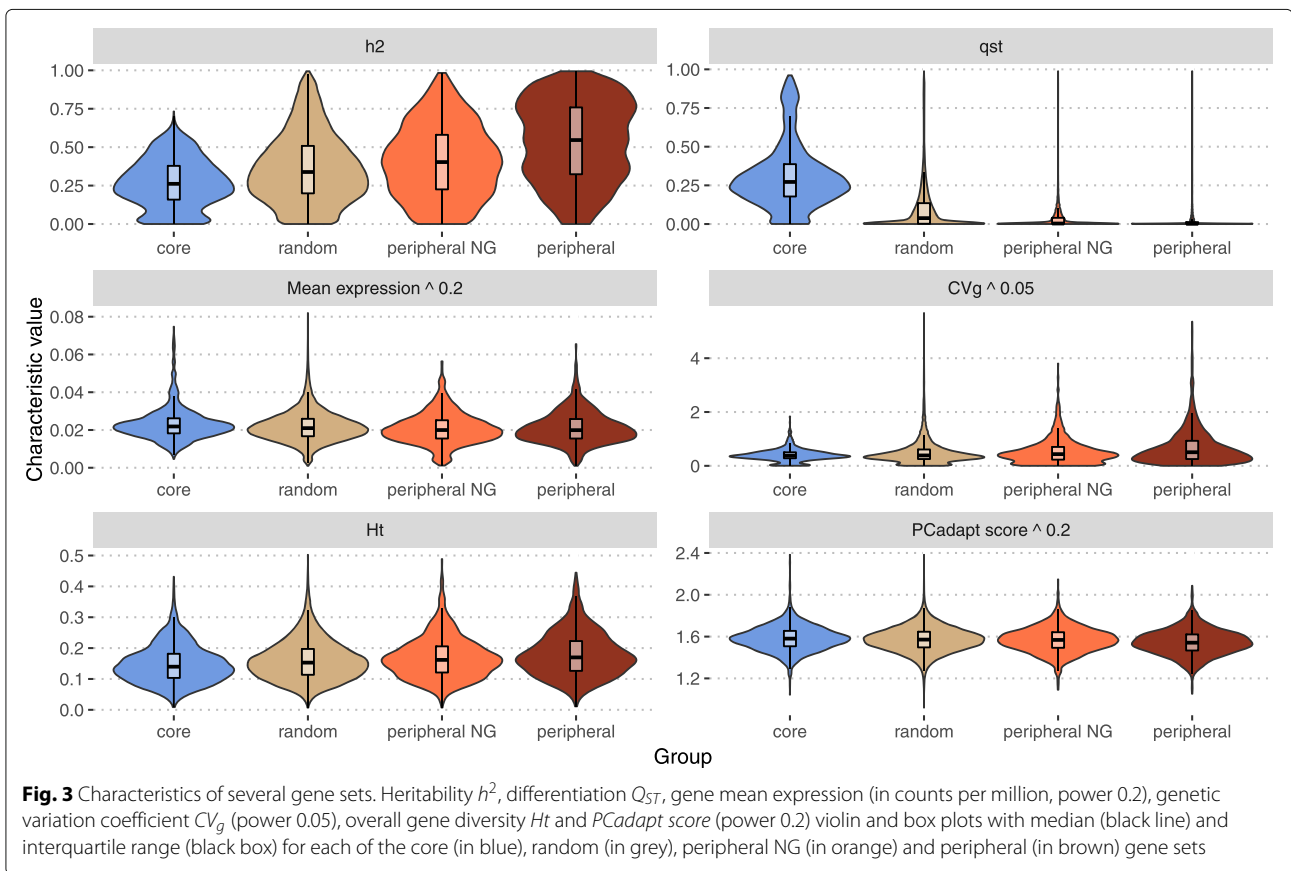
genes. For each gene, we used its highest absolute score, which corresponds to its score within the module to which it was assigned. We selected the 10% of genes with the highest global absolute scores to define the core genes group, and 10% with the lowest global absolute scores to define the peripheral genes group. Finally, we selected 100 samples of 3422 (10%) random genes as control groups (Suppl. Fig. 4, bottom panel).

One particular module from the WGCNA clustering is the grey module. This module gathers genes with low membership. In our case, it is the 2nd largest module, with 7674 genes (23% of the full set). It gathers the vast majority of genes with very low kME (Suppl. Fig. 4, bottom panel) and 99% of the peripheral genes set (Suppl. Table 2). While it is typically discarded in classic clustering studies, we chose to maintain it and rather understand its composition and role. Therefore, the peripheral gene set gathering the 10% lowest kME grey module genes was added to the comparative study. An

extra gene set was considered to complete the set of gene scenarios, one that involved low kME genes that did not belong to the grey module (subsequently called "peripheral NG", NG for "no grey").

Heritability and population differentiation of modules

To get further insights into the biological role of core and peripheral genes at population levels, we compared the distribution of various characteristics among gene sets (Fig. 3): gene expression level, several classical population statistics, including heritability (h^2), coefficient of quantitative genetic differentiation (Q_{ST}), coefficient of genetic variation (CV_g), gene diversity (Ht), and a contemporaneous equivalent to F_{ST} for genome scans ($PCadapt$ score). Gene expression level, h^2 , Q_{ST} , and CV_g were computed from gene expression data, while Ht and $PCadapt$ score [28] were computed from polymorphism data (SNP) and averaged per gene model. For more details see the "Materials and Methods" section.



Globally, there is a clear trend from core to random, to peripheral NG and to peripheral among these characteristics: with an increase for h^2 , CV_g and H_t , and a decrease for Q_{ST} , expression and $PCadapt$ score. The only differences that are not significant according to a Wilcoxon rank sum test and after Bonferroni correction are those between peripheral NG and peripheral sets in gene expression (p -value = 0.14) and between random and peripheral NG sets in the $PCadapt$ score (p -value = 0.39). All the other comparisons have p -values below 0.001.

Altogether, these statistics showed clear differences between core and peripheral genes: core genes are highly expressed, highly differentiated between populations in their expression and by their allele frequencies at linked markers, and with generally low levels of genetic variation. Contrastingly, peripheral genes are poorly expressed, poorly differentiated between populations, with generally higher genetic variation.

Boruta gene expression selection

In addition to previous gene sets building (full, core, random, peripheral NG and peripheral), we wanted to have a set of genes being relevant for their predictability of the phenotype. Our hypothesis here was that this set would be the one that enables the best prediction of a given trait

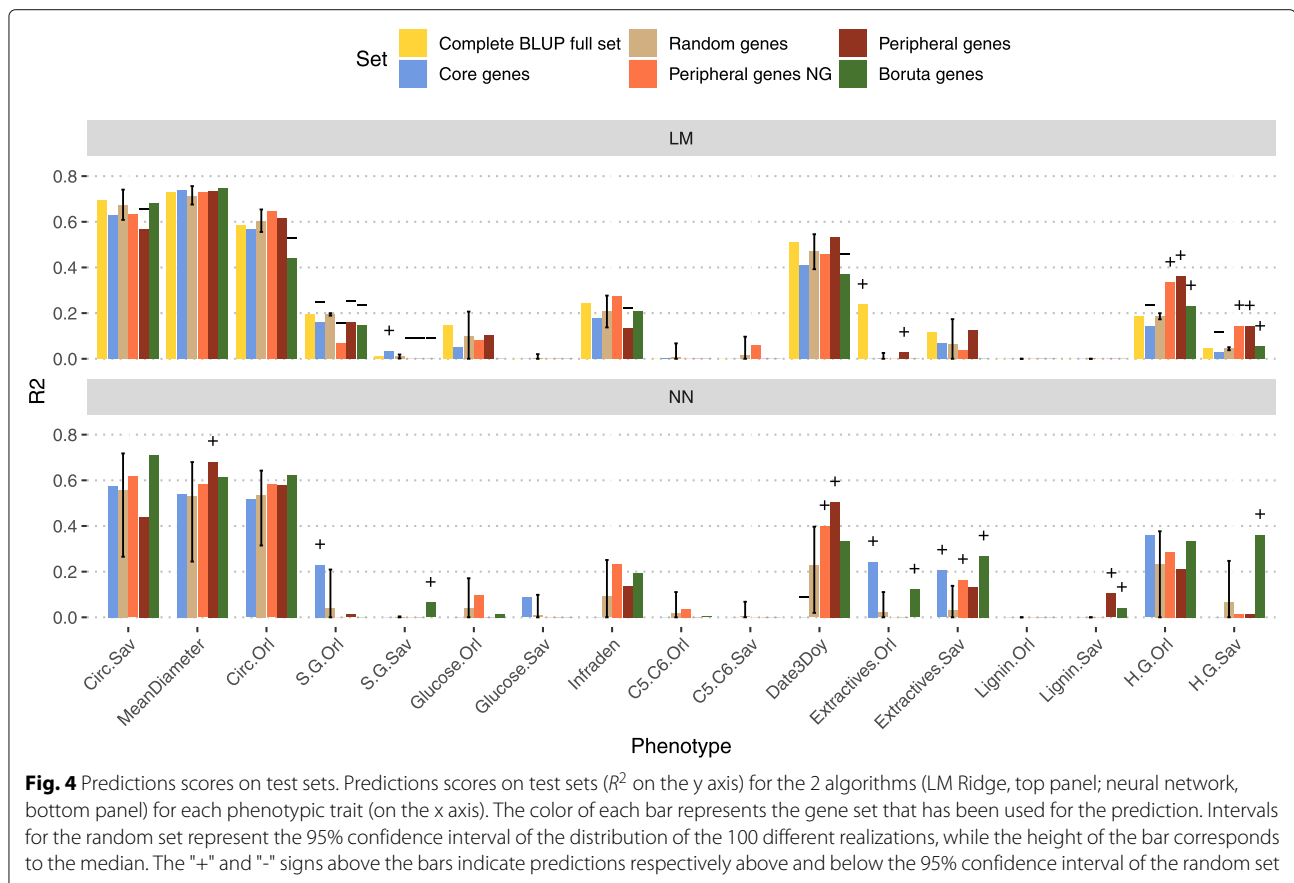
but with a limited gene number. For that purpose, we performed a Boruta (Boruta R package (v6.0.0) [29]) analysis on the full gene set with 60% of the genotypes (training set). This algorithm performs several random forests to analyze which gene expression profile is important to predict a phenotype. We tested 4 different threshold p -values for this algorithm, as we originally wanted to relax the selection and eventually get sets of different sizes. However, the number of genes selected decreased while relaxing the p -value (613, 593, 578 and 578 respectively for 0.01, 0.05, 0.1 and 0.2). Among the 4 p -values tested, 190 genes were systematically selected (114 are core, 2 are peripheral NG and 2 are peripheral genes), and 153 were selected on 3 of the 4 p -value sets (73 are core, 4 are peripheral NG and 4 are peripheral genes). By averaging across the 4 p -values tested, there was a 6.61 mean over-representation of core genes and 0.30 and 0.31 under-representation of respectively peripheral NG and peripheral genes (Suppl. Fig. 5). In the end, with a p -value of 0.01, a pool of 613 unique gene expressions was found to be important to predict our phenotypes. Traits with the highest number of important genes are related to growth. For the other traits, we always have more genes selected when the trait is measured in Orléans compared to Savigniano (respective medians of 23 and 10), which fits well

with the fact that RNA collection was performed on trees in Orléans. On average, genes that were specific to single traits represented 94% of selected genes, 1 gene was shared across sites for a given trait, genes shared by trait category (growth, phenology, physical, chemical) were 4%, and genes shared among all traits were 2%.

Phenotype prediction with gene expression

For our 6 genes sets (full, core, random, peripheral NG, peripheral and Boruta), we trained two contrasting classes of models to predict the phenotypes: an additive linear model (ridge regression, LM) and an interactive neural network model (NN). For the former, we used ridge regression to deal with the fact that for all gene sets the number of predictors was larger than the number of observations. For the latter, we chose NN as a machine-learning method, which is not subjected to dimensionality problems [30] and is able to capture interactions without a priori explicit declaration between the entries, here gene expressions. These contrasting models let us capture more efficiently either additivity or interactivity and are thus likely to inform us about the preferential mode of action of each gene set depending on their relative performances in predictability. Figure 4 shows that for LM

with ridge regression, the best gene set to predict phenotypes was on average the full set, as expected because it contains more information, followed, more surprisingly, by the peripheral and peripheral NG genes set, then the random, core and Boruta sets (respective mean prediction R^2 across all traits of 0.22, 0.21, 0.20, 0.19, 0.18 and 0.17). However, these advantages among sets were relatively small, when compared to the framework of random sets given by the 95% confidence interval from 100 realizations (95% CI, Fig. 4). Specifically, no differences were observed between random and alternative gene sets for most of the traits, with no overall set outperforming clearly the others when accounted only for traits showing significant differences with respect to 95% CI (Suppl. Fig. 6). For NN and on average terms of R^2 , random genes were the worst set, followed by core, peripheral, peripheral NG and Boruta sets (respective mean prediction R^2 across all traits of 0.14, 0.16, 0.17, 0.18 and 0.22). Again, advantages were small when compared to the reference 95% CI from random realizations. Unlike LM, however, NN yielded some net advantage for alternative sets with significant traits being mostly upwardly placed in their performances (higher R^2) with respect to the 95% CI. Among the sets with most significant cases were the



Boruta, then peripheral genes, peripheral NG genes, and Core genes. We have not been able to compute NN models with the full set as the number of predictors remains too large to be fitted with the computational power being available on computing clusters. Across phenotypes, predictions were generally slightly less variable under NN than under the ridge regression counterpart (interquartile range mean division by 1.12).

To further investigate the behavior of genes with different positions in the network with respect to the prediction model used, we computed 2 types of differences: (i) between LM and NN prediction scores for each gene set, and (ii) between core and peripheral genes sets for LM and for NN models (Suppl. Fig. 7). As a null reference for inference in the comparison between peripheral and core sets, we computed the differences between all the 100 random sets, for a total of 4950 differences corresponding to all pairwise differences, excluding reciprocals and self-comparisons. In the top panel, a positive difference indicates that LM predicted better than NN and *vice versa*, while in the bottom panel, a positive difference indicates an advantage of core genes sets over peripherals and, conversely, a negative difference indicates an advantage of peripheral genes. In any of the two panels, we did not detect any systematic difference, favouring one of the modeling options across traits or one of the gene sets across traits. Moreover, the few cases where a difference is to be noted were certainly due to very poor prediction scores. The only difference that can be noted by its magnitude is the difference between core and both peripheral genes in NN for the date of bud burst (Date3Doy), in favor of the peripheral genes.

Finally, we investigated to what extent trait h^2 or Q_{ST} would influence the prediction scores of each combination of set and algorithm. We found a positive and significant relationship between trait h^2 or Q_{ST} and prediction accuracy irrespectively of the gene set or the prediction method considered (Spearman's rank correlations ranged from 0.36 to 0.56 and from 0.27 to 0.47 for trait h^2 and Q_{ST} , respectively). When looking specifically at sets and methods, we did not find many cases showing significantly higher correlations between h^2 or Q_{ST} and prediction than the random sets. For h^2 , only the Boruta set in LM was above the 95% CI, while for Q_{ST} , only the Complete set in LM and the Boruta and Peripheral sets in NN were above the 95% CI. We further separated traits according to whether their Q_{ST} was above or below the 99th percentile of the F_{ST} . The rationale under this split is that because core genes are more differentiated between populations than random or peripheral genes, we should expect them to predict better those traits with a similar structuration behavior and *vice versa*. We found that traits above the 99th percentile of the F_{ST} were systematically better predicted than less differentiated traits. However, we did not

find significant differences between gene groups once the difference between traits was taken into account.

Discussion

Characterizing the way genes contribute to phenotypic variation could prove highly valuable to better understand the genetic architecture of complex traits. With the advent of omics data, a huge amount of information is nowadays becoming available to fill the gap between variations at the DNA and phenotype levels [31]. Such gap-filling can be obtained from multiple sources. For instance, accounting for the number of shared neighbors between two genes informs on subsequent protein-protein interactions bringing further biological meaning [32]. It is by the use of gene expression data that the present study aimed at gaining insights into the genetic architecture behind complex traits.

One key premise in the study was the availability of a common garden experiment comprising relevant samples of natural variation, in our case black poplar from Western Europe. Such an experimental setting makes it possible to accurately evaluate phenotypes to calibrate and serve as a target for predictions. Indeed, evaluating all the genotypes in a given location with experimental design and replicates enabled to unravel the confounding between genotype and macro-environment (or micro-environment) that typically occur when considering genotypes in the wild [33]. Likewise, RNA-seq data were collected on up to two biological replicates in the common garden and also corrected for environmental and design covariables, to obtain the genotypic BLUP, which is the genetic value of the genotype. Such adjustments at both phenotypic and genomic ends provided proper grounds with reasonable confidence in the absence of undesirable effects for the study of associations between the two sources of data.

Two recent works used RNA-seq in natural populations of plants to build co-expression networks and study the relationship between network topology and patterns of natural selection [3, 4]. While they found differences in natural selection among genes given their connectivity within networks, they did not investigate how these differences affect phenotypic variation. We thus embraced the commonly used WGCNA approach [5] to build the co-expression network within our dataset in order to study the relationship between gene connectivity and phenotypic prediction. This clustering of genes gave us different groups that we found to be differently correlated to traits values and according to sites. However, this method was simply for us a way to obtain a centrality or connectivity score for each gene, with the subsequent possibility to classify them into core and peripherals. The biological interpretation of correlations between gene groups and traits would clearly deserve further work which is beyond

the scope of the present study. We based our definition of core and peripheral on Mähler et al. [4], as respectively the 10% most central and most peripheral genes. The only specificity of our work here is that we did not discard, as it is classically done (called pruning in the WGCNA manual), the genes from the grey group, i.e. those showing a poor membership to any other module. We considered instead two alternative peripheral sets by keeping or excluding genes from the grey group. The pertinence of kME as a classification criterion became evident in our study when looking at the differences between core and peripherals in terms of classic quantitative and population genetic parameters. Core genes (high kME) showed high levels of population differentiation, mostly in quantitative genetic terms (Q_{ST}), while being simultaneously less variable than the rest of the genes. Such results would suggest that core genes are genes potentially subjected to divergent selection, with subsequently reduced levels of genetic variation, and involved in local adaptations. Contrarily, peripherals (low kME) showed larger levels of variation with respect to their expression level and little structure across populations, suggesting less selection pressure or weaker connection to selected traits, with mostly stabilizing selection patterns across populations. Therefore, despite the fact that a subdivision in core and peripherals is somehow an oversimplification, an extreme contrast of an otherwise continuous phenomenon, it helped to reveal the different natures of genes characterized by extreme values of kME.

To further test whether this gene categorization matters for trait prediction, we decided to go one step further by trying to predict traits from the different gene sets. We also wanted to have a gene set designed to be composed of good predictors of the traits. We thus used the Boruta algorithm [29] that performs random forest predictions by selecting the genes with the highest prediction importances. We have to keep in mind that random forest algorithms allow for implicit interactions between predictors (here gene expressions [34–36]). Results pinpointed again one feature differentiating the behavior of core and peripheral genes. Cores were largely overrepresented in the different Boruta selections (by at least 38% of Boruta genes), involving systematically the same 114 genes across all threshold p -values (or 153 over 3 values). Peripherals were systematically underrepresented to a very large extent (less than 7%). Although the remaining genes, neither cores nor peripherals according to our previous definition, were the majority (53%) among the ones selected by Boruta, they were sampled from a vaster pool of more than 27,000 genes. Another important result from the Boruta selections is the fact that relaxing the p -value threshold (from 0.01 to 0.2) did not increase the size of the resulting selection set, while the set itself could change partially in composition across different thresholds. One

can assume that relaxing the threshold would lead to increasing the number of features if these acted independently and contributed with novel information. The fact that numbers did not change substantially, while the composition was indeed impacted, leads to thinking that features are deeply interconnected and do not add up independently. This would suggest that different arrangements of genes could contain comparable levels of information or, in other words, that genes bear some redundancy through networks of interactivity.

With these 6 genes sets, we predicted 17 phenotypic traits with 2 alternative algorithms, one expected to capture mostly additivity between predictors (LM), the other one interactivity (NN). As expected, the full set resulted in best predictions with the LM model (NN not available), as it comprised all available genetic information but it rarely predicted above the random 95% CI. Furthermore, core genes were far from being the best set to predict the different traits under either of the two algorithms. Such results would be a priori surprising considering previous statements on the composition of Boruta selection where cores had an important contribution. The key difference, however, is that cores were not the only contributors to the Boruta sets. It seems that cores are able to summarize key information for quality predictions but require a complementary contribution from other interacting genes to round up the optimal set. This is better reflected by the performance of the Boruta set, which obtained the best performance predicting traits under the NN algorithm. To some extent, the NN algorithm exploits the interactivity between features (genes) already present in the Boruta set, itself obtained through the random forest heuristics that are particularly sensitive to interactions. The high connectivity of high kME value core genes is well captured by interaction sensitive algorithms to improve prediction.

In a contrasting way, the core set performed poorly under LM modeling, where the two classes of peripherals obtained the best predictabilities. Such a performance from peripherals is somehow surprising, in the sense that this class of genes, notably the grey module, is usually pruned from transcriptomic studies, while they seem nonetheless to harbor important biological information that is relevant to the trait variation. Judging from the nature of the LM modeling, peripherals would have more a type of additive gene action, which could be in turn a penalizing feature when a reduction in the number of genes operates to focus only on the most relevant ones (*i.e.* underrepresentation of peripherals in Boruta set). Thus, peripherals appear to be relevant when allowed to contribute cumulatively to prediction, although they can be otherwise easily summarized by more integrative genes when variable selection procedures operate to obtain optimal sets. It is important to note, however, that adding peripherals (following an increasing kME) beyond the

numbers present in their original sets did not improve predictability (Suppl. Fig. 8), suggesting the existence of a plateau in their capacity to explain trait variation. The low connectivity of peripheral genes, reflecting independent features, is best exploited by linear model approaches capturing mostly additive genetic actions.

Finally, random sets offered a convenient framework for inferences when comparing gene sets. Their performance in terms of predicting quality was never the best under either of the alternative modeling approaches (LM or NN) but was good enough to suggest that relevant information can be nevertheless obtained from many different gene sets, pointing at some degree of pervasive redundancy in the genetic architecture of traits. In practical terms, when a trait prediction is required but there is no biological a priori on the choice of genes, a random set modeled through LM appears like a satisfying solution. This is not far from the SNP based counterpart in genome-wide evaluation [37], where markers are often a choice that is not driven by biological context. However, if some previous selection of genes is required, the combination of Boruta selection and subsequent NN modeling has been shown here to be a good option for predictability on a reduced genic panel. Indeed, Boruta or any other NN option are advantageous alternatives in genomic evaluation for breeding to more classic methods, often based on the imposition of a priori constraints for shrinkage or variable selection [38].

One of the particularities of core genes, that of showing highly structured genetic variation among populations, led us to think that they might be preferentially involved in traits also showing high levels of Q_{ST} . Such a hypothesis was not confirmed by our results, where highly structured traits were generally better predicted than traits with no apparent structure, but with no clear differences in such an advantage between gene sets. Therefore, the highly structured core genes did not contribute to improving the prediction of highly structured traits, suggesting that trait covariation between populations is affected by other genic sources not conveniently unraveled here. It is important to note that prediction quality is highly variable between traits, somehow masking the differences that might be found between gene sets. We have already pinpointed the relevance of kME in establishing a gradient of genes whose extremes show different behaviors in quantitative and population genetics statistics. These extremes also contribute differently to the explanation of phenotypic variability, through the light of different prediction models. One aspect that remained unanswered, however, is to what extent kME is also relevant to prediction without circumscribing our scope to the extremes. When computing the correlations between connectivity (kME) and prediction coefficients (importance in terms of effect) from LM across all the full set of genes, results showed that there

are some strong positive correlations for three of the traits (Circ.Orl, S.G.Orl and Extractives.Orl). However, there is not a systematic trend across all the traits, suggesting that other differences in their genetic variability and genomic architectures might be also of importance here.

In the end, differential connectivity as reflected by our kME gradient from gene expressions pinpoints the importance of mechanisms of gene interactions in the genetic architecture of traits. On top of the DNA sequence, the superposing layer of transcriptomics adds up the intermediate pattern of gene interactions and physiological epistasis, before the final level of phenotypic expression [39]. It is important to note, however, that such gene interaction at the transcriptomic level is not directly or necessarily related to epistasis in the context of statistical genetics literature, i.e. the interaction effect between alleles from different loci on a given phenotype [40]. The extent to which connectivity or transcriptomic interactivity relates to that level of epistasis is beyond the scope of current work but clearly deserves further investigation.

Conclusion

This work shows that all genes seem important to some extent to predict phenotypes. If the Boruta selection leads us to think that core genes may be very important, prediction results across a range of phenotypes underlined that they are not the only ones. The information that those core genes contain has to be completed by other complementary genes. Likewise, on the other extreme of our networks, peripherals seem also to bear enough biological information to build up sounding predictions. Our analytical approach, by looking at the specific roles of genes with different networking connectivities, highlights the importance of the gene system as a whole in explaining phenotypic variation rather than that of particular sets of genes. Our work is globally in accordance with the recent work on the omnigenic model [41, 42], stating that all genes expressed in an organ participate in the traits of that organ. We were also able to predict phenotypes of an organ or at the organism level, with gene expression from another organ. However predicting and explaining are 2 different things, and the information beared by some genes may be too redundant to lead us to good mechanistic models, without further integration of biological information filling the gap between sequences and phenotypes.

Methods

Samples collection

As described in previous works [24, 43], we established in 2008 a partially replicated experiment with 1160 cloned genotypes, in two contrasting sites in central France (Orléans, ORL) and northern Italy (Savigliano, SAV). At ORL, the total number of genotypes was 1,098 while at

SAV there were 815 genotypes. In both sites, the genotypes were replicated 6 times in a randomized complete block design. The experiments were carried out in accordance with local legislation. At SAV, the trees were pruned at the base after one year of growth (winter 2008–2009) to remove a potential cutting effect and were subsequently evaluated for their growth and wood properties during winter 2010–2011. At ORL, the trees had the same pruning treatment after two years of growth (winter 2009–2010) and were also subsequently evaluated for growth and wood properties after two years (winter 2011–2012). After evaluation, we pruned again for a new growth cycle. In their fourth year of growth of this third cycle (2015), 241 genotypes present in two blocks of the French site were selected to perform sampling for RNA sequencing. In the end, we obtained transcriptomic data from 459 samples, 218 genotypes duplicated in the two blocks and 23 genotypes available from only one block. These 241 genotypes were representative of the natural west European range of *P. nigra* through 11 river catchments in 4 countries (Fig. 1). More details on the origin of these genotypes including their depositary are available in the [GnpIS Information System](#) [44], using the keys "Black poplar" and "POPULUS_NIGRA_RNASEQ_PANEL" for the fields "Crops" and "Germplasm list", respectively.

We described 14 of the 17 phenotypic traits in previous work [24]. Briefly, these traits can be divided into two categories, growth traits and biochemical traits which were all evaluated on up to 6 clonal replicates by genotype at each site after two years of growth in the second cycle. The first set is composed of the circumference of the tree at a 1-meter height measured in Savigliano at the end of 2009 (CIRC2009.Sav) and in Orléans at the end of 2011 (CIRC2011.Orl). The second set is composed, each time at both sites, of measures of ratios between the different components of the lignin, p-hydroxyphenyl (H), guaiacyl (G) and syringyl (S) (H.G.Orl, H.G.Sav, S.G.Orl and S.G.Sav), measures of the total lignin content (Lignin.Orl : measure of the lignin in Orléans, Lignin.Sav: measure of the lignin in Savigliano), measure of the total glucose (Glucose.Orl and Glucose.Sav), measure of ratio between 5 and 6 carbon sugars (C5.C6.Orl and C5.C6.Sav) and measure of the extractives (Extractives.Orl and Extractives.Sav). For each of these traits, we computed mean values per genotype previously adjusted for microenvironmental effects (block or spatial position in the field).

The 3 remaining traits were measured in 2015 on the trees harvested for the RNA sequencing experiment (2 replicates per genotype). They include the mean diameter of the stem section harvested for RNA sequencing (MeanDiameter), the date of bud flush of the tree in 2015 (Date3Doy) and the basic density of the wood (Infraden). Date of bud flush consisted of a prediction of the day of the year at which the apical bud of the tree was in stage

3 according to the scale defined in Dillen et al. [45]. Predictions were done with a lowess regression from discrete scores recorded at consecutive dates in the spring of 2015. Wood's basic density was measured on a piece of wood from the stem section harvested for RNA sequencing following the Technical Association of Pulp and Paper Industry (TAPPI) standard test method T 258 "Basic density and moisture content of pulpwood".

Transcriptome data generation

We sampled stem sections of approximately 80 cm long starting at 20 cm above the ground and up to 1 meter in June 2015. The bark was detached from the trunk in order to scratch young differentiating xylem and cambium tissues using a scalpel. The tissues were immediately immersed in liquid nitrogen and crudely ground before storage at -80°C pending milling and RNA extraction. Prior to RNA extraction, the samples were finely milled with a swing mill (Retsch, Germany) and tungsten beads under cryogenic conditions with liquid nitrogen during 25 seconds (frequency 25 cps/sec). About 100 mg of milled tissue was used to isolate separately total RNA from xylem and cambium of each tree with RNeasy Plant kit (Qiagen, France), according to manufacturer's recommendations. Treatment with DNase I (Qiagen, France) to ensure the elimination of genomic DNA was made during this purification step. RNA was eluted in RNase-DNase free water and quantified with a Nanodrop spectrophotometer. RNA from xylem and cambium of the same tree were pooled in an equimolar extract (250 ng/ μL) before sending it to the sequencing platform.

RNA-seq experiment was carried out at the platform POPS (transcriptOmic Platform of Institute of Plant Sciences - Paris-Saclay) thanks to IG-CNS Illumina HiSeq2000. RNA-seq libraries were constructed using TruSeq Stranded mRNA SamplePrep Guide_15031047_D protocol (Illumina®, California, U.S.A.). The RNA-seq samples have been sequenced in single-end reads (SR) with an insert library size of 260 bp and a read length of 100 bases. Images from the instruments were processed using the manufacturer's pipeline software to generate FASTQ sequence files. Ten samples by lane of HiSeq2000 using individually barcoded adapters gave approximately 20 millions of SR per sample. We mapped the reads on the *Populus trichocarpa* v3.0 transcriptome with bowtie2 v2.4.1 [46], and obtained the read counts for each of the 41,335 transcripts by homemade scripts (a median of 17 millions of reads were mapped per sample, with a minimum of 6 and a maximum of 42 million). *Populus trichocarpa* is considered the reference genome for the *Populus* genus with a high quality annotation, which is why we used it to map and quantify our data. In addition, the coding region is highly conserved between the two species and, as a result, 94% of our reads mapped

on the *Populus trichocarpa* reference. Initially, we considered using the genotype means to reduce our data volume. However, differences between replicates were not normally distributed, because of variation in gene expression due to plasticity. We thus could not summarize our data with their mean, as it would have removed this information and finally we chose to keep replicates as separate data samples.

Filtering the non-expressed genes, normalization and transformation to obtain a Gaussian distribution

We started cleaning our raw count data by removing the transcripts without at least 1 count in 10% of the individuals. From the original 41,335 genes, 7,106 were thus removed, leaving 34,229 genes. After this first filtration, we normalized the raw count data by Trimmed Mean of M-values (TMM, edgeR v3.26.4 [47]). As most features are not differentially expressed, this method takes into account the fact that the total number of reads can be strongly influenced by a low number of features. Then, we calculated the counts per millions (CPM [48]).

To make the CPM data fit a Gaussian distribution, we computed a $\log_2(n + 1)$ instead of a $\log_2(n + 0.5)$ typically used in a voom analysis [48], to avoid negative values, which are problematic for the rest of the analysis.

Computing the BLUP, heritability, and Q_{ST} while correcting the co-variables

As the sampling ran along 2 weeks, we expected environmental variables to blur the signal. To understand how our data were impacted, we ran a PCA analysis to identify the impact of each cofactor (Suppl. Fig. 1). We identified the block and the sampling date and time as cofactors with a substantial impact.

A 12k bead chip [25] provided 7,896 SNPs in our population. A genomic relationship matrix between genotypes was computed with these SNPs with LDAK [49], and further split into between (mean population kinship, \mathbf{K}_b) and within-population relationship matrices (kinship kept only for the members of the same population, all the others are equal to 0, \mathbf{K}_w). These matrices were used in a mixed linear model to compute the additive genetic variances between and within populations for the expression of each gene:

$$\mathbf{y} = \beta_0 + \mathbf{Z}_b \mathbf{b} + \mathbf{Z}_w \mathbf{w} + \epsilon \quad (1)$$

Where, \mathbf{y} is a gene expression vector across individual trees, β_0 is a vector of fixed effects (overall mean or intercept); \mathbf{b} and \mathbf{w} are respectively random effects of populations and individuals within populations, which follow normal distributions, centered around 0, of variance $\sigma_b^2 \mathbf{K}_b$ and $\sigma_w^2 \mathbf{K}_w$. σ_b and σ_w are the between and within-population variance components and \mathbf{K}_b and \mathbf{K}_w are the between and within-population kinship matrices. \mathbf{Z}_b and

\mathbf{Z}_w are known incidence matrices between and within populations, relating observations to random effects \mathbf{b} and \mathbf{w} . ϵ is the residual component of gene expression, following a normal distribution centered around 0, of variance $\sigma_\epsilon^2 \mathbf{I}$, where σ_ϵ is the residual variance and \mathbf{I} is an identity matrix.

We used the function "remlf90" from the R package breedR (v0.12.2) [27] to fit the model, with the Expectation-Maximization method followed by one round with Average-Information algorithm to compute the standard deviations. From the resulting between and within-population variance components, we computed the best linear unbiased predictors of between and within population random genetic effects ($\hat{\mathbf{b}}$ and $\hat{\mathbf{w}}$, respectively) and summed them up to obtain the total genetic value for each gene expression (BLUP). We also computed heritability (h^2) and population differentiation estimates (Q_{ST}) for each gene expression as follows:

$$h^2 = \frac{\sigma_b^2 + \sigma_w^2}{\sigma_b^2 + \sigma_w^2 + \sigma_\epsilon^2} \quad (2)$$

$$Q_{ST} = \frac{\sigma_b^2}{\sigma_b^2 + 2\sigma_w^2} \quad (3)$$

Finally, we computed for each gene expression the coefficient of genetic variation (CV_g) by dividing its total genetic variance ($\sigma_b^2 + \sigma_w^2$) by its expression mean.

Other population statistics

We further used a previously developed bioinformatics pipeline to call SNPs within our RNA sequences [50]. Briefly, this pipeline involves cleaning and quality control steps, mapping on the *P. trichocarpa* v3.0 reference genome, and SNP calling using the combination of four different callers. We ended up with a set of 874,923 SNPs having less than 50% of missing values per genotype. The missing values were further imputed with the software FImpute [51]. We validated our genotyping by RNA sequencing approach by comparing the genotype calls with genotyping previously obtained with an SNP chip on the same individuals [25]. Genotyping accuracy based on 3,841 common positions was very high, with a mean value of 0.96 and a median value of 0.99. The imputed set of SNP was then annotated using Annovar [52] in order to group the SNPs per gene model of *P. trichocarpa* reference genome. For each SNP, we computed the overall genetic diversity statistics with the hierfstat R package (v0.4.22) [53] and this statistic was then averaged by gene model in order to get information on the extent of diversity. We further computed *PCadapt* score with the *pcadapt* R package (v4.3.3) [28] with 8 retained principal components. Here again, *PCadapt* scores were then summarized (averaged) by gene-model in order to get information about their potential involvement in adaptation. Based on the

principal component analysis, *pcadapt* is more powerful to perform genome scans for selection in next-generation sequencing data than approaches based on F_{ST} outliers detection [28]. We found a positive correlation between F_{ST} and *PCadapt* score (data not shown), but *PCadapt* score highlighted differences between Core, random and peripheral gene sets (Fig. 3) while F_{ST} did not.

Hierarchical clustering

We performed a weighted correlation network analysis with the R package *WGCNA* (v1.68) [5] on our full RNA-seq gene set. We followed the recommended approach, except that we first ranked our expression data, to work subsequently with Spearman's non-parametric correlations and avoid problems due to linear modeling assumptions. We first chose the soft threshold with a power of 12, which is the recommended value for signed networks (and default value in *WGCNA*) ($R^2 = 0.81$, connectivity: mean = 195.17, median = 9.23, max = 1403.96, Fig. 2a). Then, we used the automatic module detection (function "blockwiseModules") via dynamic tree cutting with a merging threshold of 0.25, a minimum module size of 30 and bid-weight midcorrelations (Fig. 2b). All other options were left to default. This also computes module eigengenes. To sort the traits, we clustered their scaled values with the *pvclust* R packages (v2.2.0) [54], the Ward agglomerative method ("Ward.D2") on correlations (Fig. 2b, Fig. 2c, Suppl. Fig. 2). The clustering on euclidean distance results in the exact same hierarchical tree. Correlations between traits and gene expression or module eigengenes were computed as Spearman's rank correlations (Fig. 2b, c).

Machine learning

Boruta gene expression selection

In addition to the inconvenience of working with a large number of features (time and power consumption), most machine learning algorithms perform better when the number of predicting variables used is kept as low as the optimal set [55]. We thus performed an all relevant variable selection [56] with the Boruta function [29] from the eponym R package, with 4 *p*-value thresholds (1, 5, 10 and 20%), on the training subpart of the full gene expression set, for each phenotype independently. Then, features that were not rejected by the algorithm were pooled together, so that all the important genes were in the selected gene pool, one pool for each *p*-value threshold. The enrichment or depletion in core or peripheral genes in each of these pools was evaluated by Fisher's exact test for count data ("fisher.test" function in the stats R package (v3.6.3) with default parameters).

Models

Both additive linear model (ridge regression) and interactive neural network models were computed by the R

package *h2o* (v3.30.0.2) [57]. They both used the gene expression sets as predictors and one phenotypic trait at a time as a response. Datasets were split by the function "h2o.splitFrame" into 3 sets, a training set, a validation set and a test set, with the respective proportions of 60%, 20%, and 20%. We checked that the split preserves the distribution of samples within populations. The training set was used to train the models, the validation set was used to validate and improve the models, while the test set was used to compute and report prediction accuracies as R^2 between observed and predicted values within this set and using the function "R2" of the R package *compositions* (v1.40.2) [58]. This set has never been used to improve the model and therefore represents a proxy of new data, avoiding the report of results from overfitted models. All the reported predictions scores were computed on this test set. These results are thus representing real-life predictions and are not subject to over-fitting.

For linear models, we used the function "h2o.glm" with default parameters, except 2-folds cross-validation and alpha set at zero to perform a ridge regression. The same splits and score reporting methods were used.

Neural networks have the reputation to be able to predict any problem, based on the Universal approximation theorem [59, 60]. However, this capacity comes at the cost of a very large number of neurons in one layer, or a reasonable number of neurons per layer in a high number of layers. Both settings lead to difficult interpretation when very many gene expressions are involved. In that sense, we chose to keep our models simple, with two layers of a reasonable number of neurons. This obviously comes at the price of lower prediction power. However, we believe that these topologies give us the power to model 2 levels of interactions between genes (1 level per layer). Furthermore, since both methods yielded comparable prediction R^2 (median ridge regression $R^2 = 0.19$, mean neural network $R^2 = 0.173$), this complexity seemed appropriate. To find the best models for neural networks, we computed a random grid for each response. We tested the following four hyperparameters: (i) activation function ("Rectifier", "Tanh", "RectifierWithDropout" or "TanhWithDropout"); (ii) network structure; (iii) input layer dropout ratio (0 or 0.2) (iv) L1 and L2 regularization (comprised between 0 and 1×10^{-4} , with steps of 5×10^{-6}). Network structure corresponded to the number of neurons within each of the two hidden layers, which was based on the number of input genes (h). The first layer was composed of h , $\frac{2}{3}h$ or $\frac{1}{3}h$ neurons. The second layer had a number of nodes equal or lower to the first one and was also composed of h , $\frac{2}{3}h$ or $\frac{1}{3}h$ neurons. This represented a total of 6 different structures. We performed a random discrete strategy to find the best search criteria, computing a maximum of 100 models, with a stopping tolerance of 10^{-3} and 10 stopping rounds. Finally, "h2o.grid" parameters

were the following: the algorithm was "deeplearning", with 10 epochs, 2 fold cross-validation, maximum duty cycle fraction for scoring 0.025, and constraint for a squared sum of incoming weights per unit 10. All other parameters were set to default values. The best model was selected from the lowest RMSE score within the validation set.

Supplementary information

Supplementary information accompanies this paper at <https://doi.org/10.1186/s12864-020-06809-2>.

Additional file 1: Suppl. Table 1. Module membership of each gene.

Additional file 2: Suppl. Table 2. Distribution of core, peripheral and peripheral no grey genes across modules.

Additional file 3: Suppl. Fig. 1. PCA score plots on gene expression data. Each plot represents the distribution of the individuals on the 2 first axes of the PCA (representing 17,7% of the variation), colored by class of various experimental factors (Xylem and cambium scraper, extractor and extraction method, population, sequencing column, line and plate, the growth rate at harvest, sampling date, time, temperature, solar radiation, humidity and wind speed). Cofactors related to weather are presented in the 6 lower plots.

Additional file 4: Suppl. Fig. 2. Traits hierarchical ascendant clustering dendrogram. Clustering was performed from the correlations between traits with Ward method ("Ward.D2") by the R package pvclust. Approximately Unbiased (au, in red) and Bootstrap Probability (bp, in green) p -values indicated the degree of belief associated with clusters. Highly supported modules are framed by a red square, grouping (a) the mean sample diameter with the two circumference traits, (b) the S/G ratios with glucose composition, (c) the two C5/C6 together, and (d) the H/G ratios.

Additional file 5: Suppl. Fig. 3. Relationship between Spearman's correlations between module-trait (y-axis) and gene significance-kME (x-axis).

Additional file 6: Suppl. Fig. 4. Histograms of the centrality scores without (top) or with (bottom) the grey group. Core, peripheral and peripheral without grey sets are represented respectively by the blue, dark orange and orange bars. Random sets are distributed across the histogram and do not appear on this figure. Distribution of genes clustered in the grey module is represented by the grey bars, white bars are for other genes.

Additional file 7: Suppl. Fig. 5. Histograms of the centrality scores for the genes selected by Boruta at different p -values thresholds. Repartition of selected genes within the following gene sets is highlighted, with core in blue, peripheral NG in orange, peripheral in brown and other (NA) in black. Four p -value thresholds for Boruta selections were considered: 0.01, 0.05, 0.1 and 0.2.

Additional file 8: Suppl. Fig. 6. Proportion of linear model (LM, top row) and neural network (NN, bottom row) predictions with a R^2 above (left column) or below (right column) the 95% confidence interval computed from the predictions with the random sets of genes for each gene set (there is no neural network model computed for the Complete BLUP full set).

Additional file 9: Suppl. Fig. 7. Difference of prediction scores between algorithms (top) and sets (bottom). On the top panel, the difference between LM and NN prediction scores for the core (in blue), random (in grey), peripheral (in brown), peripheral (in orange) and Boruta gene sets (in green). On the bottom panel, the LM differences are in red and the NN differences in turquoise and the color filling the bar represents the difference between core and peripheral genes in brown, core and peripheral NG in orange and between the random sets in grey. For the random pairs, error bars represent the first and third quartiles of the differences between pairs of randomized sets and the bar corresponds to the median.

Additional file 10: Suppl. Fig. 8. Predictions scores on test sets for increasing numbers of the peripheral genes. Violin and boxplots of prediction R^2 for the LM Ridge algorithm and for increasing sizes of the peripheral genes set (in brown) and the peripheral NG genes set (in orange), used for the predictions (in percent of the full set).

Abbreviations

CPM: Counts per million CVg: Genetic variation coefficient DNA: Deoxyribonucleic acid eQTL: Expression QTL F_{ST} : Fixation index GWAS: Genome-wide association study h^2 : Heritability H_t : Overall genetic diversity statistic kME: A measure of centrality in a module LM: Linear model NN: Neural networks model ORL: Orléans common garden experiment PCA: Principal component analysis PCadapt score: The score from analysis with PCadapt R package Q_{ST} : Coefficient of quantitative genetic differentiation QTL: Quantitative trait loci RNA: Ribonucleic acid RNA-seq: RNA sequencing SAV: Savigliano common garden experiment SNP: Single nucleotide polymorphism SR: Single-end read TMM: Trimmed mean of m-values WGCNA: Weighted gene correlation network analysis

Acknowledgements

The authors gratefully acknowledge the staff of the INRA GBFOR experimental unit for the establishment and management of the poplar experimental design in Orléans, the collection of wood samples in each site, and their contribution to phenotypic measurements on poplars in Orléans; Alasia Franco Vivai staff for management of the poplar experimental plantation in Savigliano, and M. Sabatti and F. Fabbri for their contribution to phenotypic measurements on poplars in Savigliano. We acknowledge the staff of BioForA for their contribution to RNA collection in the field. We are grateful to the genotoul bioinformatics platform Toulouse Midi-Pyrénées for providing computing and storage resources. We would also like to thank M. Nordborg for useful discussions on this work and J. Salse for useful comments on the manuscript.

Authors' contributions

AC, LS, and VS designed the experiment, discussed the results and wrote this manuscript. AC ran the in silico experiment. MCL, VB, CPL, LT, MLMM, and VS contributed the RNA-seq data production and analysis. VJ, OR and VS contributed to the SNP data production and analysis. MLMM and JCL contributed to the discussion on the methodology employed. All the authors have read and approved this manuscript.

Funding

Establishment and management of the experimental sites were carried out with financial support from the NOVELTREE project (EU-FP7-211868). RNA collection, extraction, and sequencing were supported by the SYBIOPOP project (ANR-13-JSV6-0001) funded by the French National Research Agency (ANR). The platform POPS benefits from the support of the LabEx Saclay Plant Sciences-SPS (ANR-10-LABX-0040-SPS).

Availability of data and materials

This RNA-seq project has been submitted to the international repository Gene Expression Omnibus (GEO) from NCBI (accession number: GSE128482). All steps of the experiment, from growth conditions to bioinformatic analyses are detailed in CATdb [61] according to the MINSEQE "minimum information about a high-throughput sequencing experiment". Raw sequences (FASTQ) are being deposited in the Sequence Read Archive (SRA) from NCBI (accession number: SRP188754). Information on the studied genotypes is available in the GnpIS Information System [44], using the keys "Black poplar" and "POPULUS_NIGRA_RNASEQ_PANEL" for the fields "Crops" and "Germplasm list", respectively. The code used throughout the study has been deposited in the following repository: <https://forgemia.inra.fr/aurelien.chateigner/sybiopop-code.git>.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

¹BioForA, INRAE, ONF, Orléans, France. ²BIOGECO, INRAE, Univ. Bordeaux, Cestas, France. ³Institute of Plant Sciences Paris-Saclay (IP2S), CNRS, INRAE, Université Paris-Sud, Université d'Evry, Université Paris-Saclay, Gif sur Yvette, France. ⁴Institute of Plant Sciences Paris-Saclay (IP2S), CNRS, INRAE, Université Paris-Diderot, Sorbonne Paris-Cité, Gif sur Yvette, France. ⁵MIA-Paris, AgroParisTech, INRAE, Paris, France. ⁶AGAP, Université Montpellier, CIRAD, INRAE, Montpellier SupAgro, Montpellier, France.

Received: 10 December 2019 Accepted: 8 June 2020

Published online: 22 June 2020

References

- Mackay TFC, Stone EA, Ayroles JF. The genetics of quantitative traits: challenges and prospects. *Nat Rev Genet.* 2009;10(8):565–77. <https://doi.org/10.1038/nrg2612>.
- Han Y, Gao S, Muegge K, Zhang W, Zhou B. Advanced Applications of RNA Sequencing and Challenges. *Bioinforma Biol Insights.* 2015;9s1:28991. <https://doi.org/10.4137/BBI.S28991>. NIHMS150003.
- Josephs EB, Wright SI, Stinchcombe JR, Schoen DJ. The Relationship between Selection, Network Connectivity, and Regulatory Variation within a Population of *Capsella grandiflora*. *Genome Biol Evol.* 2017;9(4):1099–109. <https://doi.org/10.1093/gbe/evx068>.
- Mähler N, Wang J, Terebieniec BK, Ingvarsson PK, Street NR, Hvidsten TR. Gene co-expression network connectivity is an important determinant of selective constraint. *PLoS Genet.* 2017;13(4):1006402. <https://doi.org/10.1371/journal.pgen.1006402>.
- Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics.* 2008;9(1):559. <https://doi.org/10.1186/1471-2105-9-559>.
- Williamson SH, Hernandez R, Fedel-Alon A, Zhu L, Nielsen R, Bustamante CD. Simultaneous inference of selection and population growth from patterns of variation in the human genome. *Proc Natl Acad Sci.* 2005;102(22):7882–7. <https://doi.org/10.1073/pnas.0502300102>.
- Josephs E, Lee YW, Stinchcombe JR, Wright SI. Association mapping reveals the role of purifying selection in the maintenance of genomic variation in gene expression. *PNAS.* 2015;112(50):1–6. <https://doi.org/10.1101/015743>.
- Sicard A, Kappel C, Josephs EB, Lee YW, Marona C, Stinchcombe JR, Wright SI, Lenhard M. Divergent sorting of a balanced ancestral polymorphism underlies the establishment of gene-flow barriers in *Capsella*. *Nat Commun.* 2015;6(1):7960. <https://doi.org/10.1038/ncomms8960>.
- Han M, Qin S, Song X, Li Y, Jin P, Chen L, Ma F. Evolutionary rate patterns of genes involved in the *Drosophila* Toll and Imd signaling pathway. *BMC Evol Biol.* 2013;13(1):245. <https://doi.org/10.1186/1471-2148-13-245>.
- Lu Y. Evolutionary Rate Variation in Anthocyanin Pathway Genes. *Mol Biol Evol.* 2003;20(11):1844–53. <https://doi.org/10.1093/molbev/msg197>.
- Rauscher MD, Lu Y, Meyer K. Variation in Constraint Versus Positive Selection as an Explanation for Evolutionary Rate Variation Among Anthocyanin Genes. *J Mol Evol.* 2008;67(2):137–44. <https://doi.org/10.1007/s00239-008-9105-5>.
- Rauscher MD, Miller RE, Tiffin P. Patterns of evolutionary rate variation among genes of the anthocyanin biosynthetic pathway. *Mol Biol Evol.* 1999;16(2):266–74. <https://doi.org/10.1093/oxfordjournals.molbev.a026108>.
- Riley RM, Jin W, Gibson G. Contrasting selection pressures on components of the Ras-mediated signal transduction pathway in *Drosophila*. *Mol Ecol.* 2003;12(5):1315–23. <https://doi.org/10.1046/j.1365-294X.2003.01741.x>.
- Yu H-S, Shen Y-H, Yuan G-X, Hu Y-G, Xu H-E, Xiang Z-H, Zhang Z. Evidence of Selection at Melanin Synthesis Pathway Loci during Silkworm Domestication. *Mol Biol Evol.* 2011;28(6):1785–99. <https://doi.org/10.1093/molbev/msr002>.
- Jovelin R, Phillips PC. Expression Level Drives the Pattern of Selective Constraints along the Insulin/Tor Signal Transduction Pathway in *Caenorhabditis*. *Genome Biol Evol.* 2011;3:715–22. <https://doi.org/10.1093/gbe/evr071>.
- Song X, Jin P, Qin S, Chen L, Ma F. The Evolution and Origin of Animal Toll-Like Receptor Signaling Pathway Revealed by Network-Level Molecular Evolutionary Analyses. *PLoS ONE.* 2012;7(12):51657. <https://doi.org/10.1371/journal.pone.0051657>.
- Wu X, Chi X, Wang P, Zheng D, Ding R, Li Y. The evolutionary rate variation among genes of HOG-signaling pathway in yeast genomes. *Biol Direct.* 2010;5(1):46. <https://doi.org/10.1186/1745-6150-5-46>.
- Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH. Why highly expressed proteins evolve slowly. *Proc Natl Acad Sci.* 2005;102(40):14338–43. <https://doi.org/10.1073/pnas.0504070102>.
- Duret L, Mouchiroud D. Determinants of Substitution Rates in Mammalian Genes: Expression Pattern Affects Selection Intensity but Not Mutation Rate. *Mol Biol Evol.* 2000;17(1):68–70. <https://doi.org/10.1093/oxfordjournals.molbev.a026239>.
- Pál C, Papp B, Hurst LD. Highly expressed genes in yeast evolve slowly. *Genetics.* 2001;158(2):927–31.
- Montanucci L, Laayouni H, Dall'Olio GM, Bertranpetit J. Molecular Evolution and Network-Level Analysis of the N-Glycosylation Metabolic Pathway Across Primates. *Mol Biol Evol.* 2011;28(1):813–23. <https://doi.org/10.1093/molbev/msq259>.
- Bloom JD, Adami C. Evolutionary rate depends on number of protein-protein interactions independently of gene expression level: response. *BMC Evol Biol.* 2004;4(1):14. <https://doi.org/10.1186/1471-2148-4-14>.
- Fraser HB, Hirsh AE. Evolutionary rate depends on number of protein-protein interactions independently of gene expression level. *BMC Evol Biol.* 2004;4(1):13. <https://doi.org/10.1186/1471-2148-4-13>.
- Gebreselassie MN, Ader K, Boizot N, Millier F, Charpentier J-PP, Alves A, Simões R, Rodrigues JC, Bodineau G, Fabbri F, Sabatti M, Bastien C, Segura V. Near-infrared spectroscopy enables the genetic analysis of chemical properties in a large set of wood samples from *Populus nigra* (L.) natural populations. *Ind Crops Prod.* 2017;107(January):159–71. <https://doi.org/10.1016/j.indcrop.2017.05.013>.
- Faivre-Rampant P, Zaina G, Jorge V, Giacomello S, Segura V, Scalabrini S, Guérin V, De Paoli E, Aluome C, Viger M, Cattonaro F, Payne A, PaulStephenRaj P, Le Paslier MC, Berard A, Allwright MR, Villar M, Taylor G, Bastien C, Morgante M. New resources for genetic studies in *Populus nigra*: genome-wide SNP discovery and development of a 12k Infinium array. *Mol Ecol Resour.* 2016;16(4):1023–36. <https://doi.org/10.1111/1755-0998.12513>.
- R Core Team. R: A Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing; 2020. <https://www.R-project.org/>.
- Muñoz F, Sanchez L. breedR: Statistical Methods for Forest Genetic Resources Analysts. 2017. <https://github.com/famuvie/breedR>. R package version 0.12-2.
- Luu K, Bazin E, Blum MGB. pcadapt: an R package to perform genome scans for selection based on principal component analysis. *Mol Ecol Res.* 2017;17(1):67–77. <https://doi.org/10.1111/1755-0998.12592>.
- Kursa MB, Rudnicki WR. Feature Selection with the Boruta Package. *J Stat Softw.* 2010;36(11):1–13. <https://doi.org/10.18637/jss.v036.i11>.
- González-Recio O, Rosa GJM, Gianola D. Machine learning methods and predictive ability metrics for genome-wide prediction of complex traits. *Livest Sci.* 2014;166:217–31. <https://doi.org/10.1016/j.livsci.2014.05.036>.
- Barbeira AN, Dickinson SP, Bonazzola R, Zheng J, Wheeler HE, Torres JM, Torstenson ES, Shah KP, Garcia T, Edwards TL, Stahl EA, Huckins LM, Nicolae DL, Cox NJ, Im HK. Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nat Commun.* 2018;9(1):1825. <https://doi.org/10.1038/s41467-018-03621-1>.
- Tong AHY. Global Mapping of the Yeast Genetic Interaction Network. *Science.* 2004;303(5659):808–13. <https://doi.org/10.1126/science.1091317>.
- de Villemereuil P, Gaggiotti OE, Mouterde M, Till-Bottraud I. Common garden experiments in the genomic era: new perspectives and opportunities. *Heredity.* 2016;116(3):249–54. <https://doi.org/10.1038/hdy.2015.93>.
- McKinney BA, Reif DM, Ritchie MD, Moore JH. Machine learning for detecting gene-gene interactions: a review. *Appl Bioinforma.* 2006;5(2):77–88. <https://doi.org/10.2165/00822942-200605020-00002>.
- Chen X, Liu CT, Zhang M, Zhang H. A forest-based approach to identifying gene and gene-gene interactions. *Proc Natl Acad Sci U S A.* 2007;104(49):19199–203. <https://doi.org/10.1073/pnas.0709868104>.
- Jiang R, Tang W, Wu X, Fu W. A random forest approach to the detection of epistatic interactions in case-control studies. *BMC Bioinformatics.* 2009;10:65. <https://doi.org/10.1186/1471-2105-10-S1-S65>.

- <https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-10-S1-S65>.
37. Meuwissen THE, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. *Genetics*. 2001;157(4):1819–29. <https://doi.org/10.1186/1471-2105-10-S1-S65>.
 38. de los Campos G, Hickey JM, Pong-Wong R, Daetwyler HD, Calus MPL. Whole-Genome Regression and Prediction Methods Applied to Plant and Animal Breeding. *Genetics*. 2013;193(2):327–45. <https://doi.org/10.1534/genetics.112.143313>.
 39. Schrag TA, Westhues M, Schipprack W, Seifert F, Thiemann A, Scholten S, Melchinger AE. Beyond Genomic Prediction: Combining Different Types of omics Data Can Improve Prediction of Hybrid Performance in Maize. *Genetics*. 2018;208(4):1373–85. <https://doi.org/10.1534/genetics.117.300374>.
 40. Cordell HJ. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Hum Mol Genet*. 2002;11(20):2463–68. <https://doi.org/10.1093/hmg/11.20.2463>.
 41. Boyle EA, Li Yi, Pritchard JK. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell*. 2017;169(7):1177–86. <https://doi.org/10.1016/j.cell.2017.05.038>.
 42. Liu X, Li Yi, Pritchard JK. Trans Effects on Gene Expression Can Drive Omnigenic Inheritance. *Cell*. 2019;177(4):1022–346. <https://doi.org/10.1016/j.cell.2019.04.014>.
 43. Guet J, Fabbri F, Fichot R, Sabatti M, Bastien C, Brignolas F. Genetic variation for leaf morphology, leaf structure and leaf carbon isotope discrimination in European populations of black poplar (*Populus nigra* L.). *Tree Physiol*. 2015;35(8):850–63. <https://doi.org/10.1093/treephys/tpv056>.
 44. Steinbach D, Alaux M, Amselem J, Choisine N, Durand S, Flores R, Keliet A-O, Kimmel E, Lapalu N, Luyten I, Michotey C, Mohellibi N, Pommier C, Reboux S, Valdenaire D, Verdelet D, Quesneville H. GnpI: an information system to integrate genetic and genomic data from plants and fungi. *Database*. 2013;2013(0):058. <https://doi.org/10.1093/database/bat058>.
 45. Dillen SY, Marron N, Sabatti M, Ceulemans R, Bastien C. Relationships among productivity determinants in two hybrid poplar families grown during three years at two contrasting sites. *Tree Physiol*. 2009;29(8):975–87. <https://doi.org/10.1093/treephys/tpp036>.
 46. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods*. 2012;9(4):357–59. <https://doi.org/10.1038/nmeth.1923>.
 47. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol*. 2010;11(3):25. <https://doi.org/10.1186/gb-2010-11-3-r25>.
 48. Law CW, Chen Y, Shi W, Smyth GK. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol*. 2014;15(2):29. <https://doi.org/10.1186/gb-2014-15-2-r29>.
 49. Speed D, Hemani G, Johnson MR, Balding DJ. Improved heritability estimation from genome-wide SNPs. *Am J Hum Genet*. 2012;91(6):1011–21. <https://doi.org/10.1016/j.ajhg.2012.10.010>.
 50. Rogier O, Chateigner A, Amanzougarene S, Lesage-Descauses M-C, Balzergue S, Brunaud V, Caius J, Soubigou-Taconnat L, Jorge V, Segura V. Accuracy of RNAseq based SNP discovery and genotyping in *Populus nigra*. *BMC Genomics*. 2018;19(1):909. <https://doi.org/10.1186/s12864-018-5239-z>.
 51. Sargolzaei M, Chesnais JP, Schenkel FS. A new approach for efficient genotype imputation using information from relatives. *BMC Genomics*. 2014;15(1):. <https://doi.org/10.1186/1471-2164-15-478>.
 52. Wang K, Li M, Hakonarson H. ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*. 2010;38(16):. <https://doi.org/10.1093/nar/gkq603>.
 53. Goudet J, Jombart T. Hierfstat: Estimation and Tests of Hierarchical F-Statistics. 2015. <https://CRAN.R-project.org/package=hierfstat>. R package version 0.04-22.
 54. Suzuki R, Shimodaira H. Pvcust: Hierarchical Clustering with *p*-Values Via Multiscale Bootstrap Resampling. 2015. <https://CRAN.R-project.org/package=pvcust>. R package version 2.0-0.
 55. Kohavi R, John GH. Wrappers for feature subset selection. *Artif Intell*. 1997;97(1-2):273–324. [https://doi.org/10.1016/S0004-3702\(97\)00043-X](https://doi.org/10.1016/S0004-3702(97)00043-X).
 56. Nilsson R, PeñaPe JM, Jmp P, Björkegren JOHANBJORKEGREN J, Tegnér JESPERT J. Consistent Feature Selection for Pattern Recognition in Polynomial Time. Technical report. 2007. http://compmed.se/files/6914/2107/3475/pub_2007_5.pdf.
 57. LeDell E, Gill N, Aiello S, Fu A, Candel A, Click C, Kraljevic T, Nykodym T, Aboyou P, Kurka M, Malohlava M. H2o: R Interface for 'H2O'. 2019. <https://CRAN.R-project.org/package=h2o>. R package version 3.22.1.1.
 58. van den Boogaart KG, Tolosana-Delgado R, Bren M. Compositions: Compositional Data Analysis. 2018. <https://CRAN.R-project.org/package=compositions>. R package version 1.40-2.
 59. Cybenko G. Approximation by superpositions of a sigmoidal function. *Math Control Signals Syst*. 1989;2(4):303–14. <https://doi.org/10.1007/BF02551274>.
 60. Hornik K, Stinchcombe M, White H. Multilayer feedforward networks are universal approximators. *Neural Netw*. 1989;2(5):359–66. [https://doi.org/10.1016/0893-6080\(89\)90020-8](https://doi.org/10.1016/0893-6080(89)90020-8).
 61. Gagnot S, Tamby J-P, Martin-Magniette M-L, Bitton F, Taconnat L, Balzergue S, Aubourg S, Renou J-P, Lecharny A, Brunaud V. CATdb: a public access to Arabidopsis transcriptome data from the URGV-CATMA platform. *Nucleic Acids Res*. 2007;36(Database):986–90. <https://doi.org/10.1093/nar/gkm757>.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions



